ADVANCEMENTS IN TECHNOLOGY-BASED ASSESSMENT: EMERGING ITEM FORMATS, TEST DESIGNS, AND DATA SOURCES

EDITED BY: Fank Goldhammer, Ronny Scherer and Samuel Greiff

PUBLISHED IN: Frontiers in Psychology and Frontiers in Education

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88963-506-1
DOI 10.3389/978-2-88963-506-1

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

ADVANCEMENTS IN TECHNOLOGY-BASED ASSESSMENT: EMERGING ITEM FORMATS, TEST DESIGNS, AND DATA SOURCES

Topic Editors:
Frank Goldhammer, Leibniz Institute for Research and Information in Education (DIPF), Germany
Ronny Scherer, Centre for Educational Measurement, Faculty of Educational Sciences, University of Oslo, Norway
Samuel Greiff, University of Luxembourg, Luxembourg

Citation: Goldhammer, F., Scherer, R., Greiff, S., eds. (2020). Advancements in Technology-Based Assessment: Emerging Item Formats, Test Designs, and Data Sources. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88963-506-1

Table of Contents

EDITORIAL
published: 20 January 2020
doi: 10.3389/fpsyg.2019.03047

Editorial: Advancements in Technology-Based Assessment: Emerging Item Formats, Test Designs, and Data Sources

Frank Goldhammer[1,2]*, *Ronny Scherer*[3] *and Samuel Greiff*[4]

[1] Educational Quality and Evaluation, DIPF - Leibniz Institute for Research and Information in Education, Frankfurt, Germany, [2] Centre for International Student Assessment (ZIB), Frankfurt, Germany, [3] Centre for Educational Measurement (CEMO), University of Oslo, Oslo, Norway, [4] Cognitive Science & Assessment, University of Luxembourg, Esch-sur-Alzette, Luxembourg

Keywords: technology-based assessment, item design, test design, automatic scoring, process data, assessment of/for learning

Editorial on the Research Topic

Advancements in Technology-Based Assessment: Emerging Item Formats, Test Designs, and Data Sources

OPEN ACCESS

Edited and reviewed by:
Yenchun Jim Wu,
National Taiwan Normal University, Taiwan

****Correspondence:***
Frank Goldhammer
goldhammer@dipf.de

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *19 December 2019*
Accepted: *23 December 2019*
Published: *20 January 2020*

Citation:
Goldhammer F, Scherer R and Greiff S (2020) Editorial: Advancements in Technology-Based Assessment: Emerging Item Formats, Test Designs, and Data Sources. Front. Psychol. 10:3047. doi: 10.3389/fpsyg.2019.03047

Technology has become an indispensable tool for educational and psychological assessment in today's world. Individual researchers and large-scale assessment programs alike are increasingly using digital technology (e.g., laptops, tablets, and smartphones) to collect behavioral data beyond the mere correctness of item responses. Along these lines, technology innovates and enhances assessments in terms of item and test design, methods of test delivery, data collection and analysis, and the reporting of test results.

The aim of this Research Topic is to present recent developments in technology-based assessment and in the advancements of knowledge associated with it. Our focus is on cognitive assessments, including the measurement of abilities, competences, knowledge, and skills, but also includes non-cognitive aspects of assessment (Rausch et al.; Simmering et al.). In the area of (cognitive) assessments, the innovations driven by technology are manifold, and the topics covered in this collection are, accordingly, wide and comprehensive: Digital assessments facilitate the creation of new types of stimuli and response formats that were out of reach for assessments using paper; for instance, interactive simulations may include multimedia elements, as well as virtual or augmented realities (Cipresso et al.; de-Juan-Ripoll et al.). These types of assessments also allow for the widening of the construct coverage in an assessment; for instance, through stimulating and making visible certain problem-solving strategies that represent new forms of problem solving (Han et al.; Kroeze et al.). Moreover, technology allows for the automated generation of items based on specific item models (Shin et al.). Such items can be assembled into tests in a more flexible way than what is possible in paper-and-pencil tests and can even be created on the fly; for instance, tailoring item difficulty to individual ability (adaptive testing) while assuring that multiple content constraints are met (Born et al.; Zhang et al.). As a requirement for adaptive testing, or to lower the burden of raters who code item responses manually, computers enable the automatic scoring of constructed responses; for instance, text responses can be coded automatically by using natural language processing and text mining (He et al.; Horbach and Zesch).

Technology-based assessments provide not only *response data* (e.g., correct vs. incorrect responses) but also *process data* (e.g., frequencies and sequences of test-taking strategies, including navigation behavior) that reflect the course of solving a test item and gives information on the

TABLE 1 | Overview of the papers.

References	Area(s) of advancement	Data types	Statistical approach	Assessment purpose (of/for learning)	Assessment domains	Key finding and advancement
Focus on new data types and sources						
Blaauw et al.	Computerized assessment of learning with multiple informants	Survey responses, platform user data	Descriptive approach	For	Vocational education	Multi-informant time-series data can inform the success of educational interventions to support students at risk
De Boeck and Scalise	Log-file and performance data to assess ColPS	Actions, response times, correctness of item responses	Confirmatory factor analysis	Of	Collaborative problem solving (PISA 2015)	Dependencies among action, time-on task, and performance indicators do not only exist at the construct but also the item (residual) level
Lindner et al.	Time-on task to identify rapid guessing	Correctness of item responses, response times	Latent class analysis	Of	Science achievement	Response times can provide information about rapid-guessing behavior and its relations to cognitive resources and test-taking effort
Naumann	Time-on task data of reading	Correctness of item responses, response times	Linear mixed modeling	Of	Reading literacy (PISA 2009)	Response times can help identify relations between item difficulties, strategic knowledge, skills, and motivation to ultimately craft a validity argument
Simmering et al.	Assessment of non-cognitive skills	Continuous process data (e.g., behavioral, physiological)	–	–	Non-cognitive skills	Challenges and limitations in using technology-enhanced assessments require consideration
von Davier et al.	Data paradigms for educational learning and assessment systems	Response behavior, test content, instructional content	e.g., machine learning	Of/For	Divers	The concept of the "data cube" can be used to label, collect and store data
Focus on innovative item designs						
Arieli-Attali et al.	Learning design	Learners' responses and use of learning support	e.g., hidden Markov modeling	For	Divers	The traditional evidence centered design models can be expanded to assess learning
Cipresso et al.	Assessment of unilateral spatial neglect	Correctness of item responses	–	–	Unilateral spatial neglect	Complex 3D environments on mobile devices are promising for the ecological assessment of unilateral spatial neglect
de-Juan-Ripoll et al.	Assessment of risk taking	Behavioral and physiological responses	–	–	Risk taking	Virtual realities (VR) can be employed to simulate hazardous situations realistically
den Ouden et al.	Computerized dynamic assessment of text comprehension skills	Correctness of item responses	Linear modeling and MTMM	For	Text comprehension	Computer-based dynamic assessments bear the potential to support students in acquiring reading skills
Horbach and Zesch	Automated content scoring	Written text	Machine learning	Of	Diverse	Automated content scoring approaches can take into account the variance in learner answers
Kroeze et al.	Automated feedback generation	Written text, actions, correctness of item responses	Descriptive approach, linear model	Of/For	Scientific inquiry in economics and physics	Automated feedback on scientific hypotheses can agree with human ratings to a great extent, and students who receive it are likely to develop better hypotheses than those who don't

(Continued)

TABLE 1 | Continued

References	Area(s) of advancement	Data types	Statistical approach	Assessment purpose (of/for learning)	Assessment domains	Key finding and advancement
Focus on innovative test designs						
Born et al.	Computerized adaptive testing and test equating	Correctness of item responses	Item response theory	Of	–	Equating designs and CAT can be combined through a continuous calibration strategy
Csapó and Molnár	Assessment for teaching and learning	Correctness of item responses	Item response theory	For	Mathematics, science, and reading	Teaching and learning can be supported on a large scale by online assessment solutions (authoring, assembly, scoring, delivery, feedback)
Molnár and Csapó	Computerized assessment of cognitive development	Correctness of item responses	Confirmatory factor analysis and structural equation models	Of/For	Mathematics, science, and reading competence	Computerized assessments can capture differences in the academic performance on tests in mathematics, science, and reading across grade levels and make visible the psychological dimension of learning
Rausch et al.	Embedded experience sampling for assessing non-cognitive skills	Survey responses, correctness of item responses	MTMM, item response theory	Of	Non-cognitive facets of problem solving	Embedded experience sampling provides an approach to assess non-cognitive facets of competences through multiple self-reports
Zhang et al.	Computerized adaptive testing of Internet addiction	Survey responses	Item response theory	Of	Internet addiction	A computerized adaptive test of Internet addiction assessed the construct accurately and efficiently, and provided evidence for both the reliability and validity of the resultant test scores
Focus on statistical approaches						
Han et al.	Data mining using random forests to predict item performance	Actions, response times, correctness of item responses	Tree-based model	Of	Problem solving (PISA 2012)	A random forest algorithm can generate and select features from the process data that predict students' item responses
He et al.	Text mining and item response data to identify PTSD	Written text, survey responses	Item response theory and text classification	–	Post-traumatic stress disorder	Combining text classification and item response theory models provides an efficient approach to estimating the latent trait
Shin et al.	Topic modeling for item distractor generation	Written text	Machine learning	Of	Knowledge and skills in biology	Latent topic modeling supports the identification of students' misconceptions in biology and aids the development of distractors

path toward the solution (Han et al.). Process data, among others, have been used successfully to evaluate and explain data quality (Lindner et al.), to define process-oriented latent variables (De Boeck and Scalise), to improve measurement precision, and to address substantial research questions (Naumann). Large-scale result and process data also call for data-driven computational approaches in addition to traditional psychometrics and new concepts for storing and managing data (von Davier et al.).

The contributions of this Research Topic address how technology can further improve and enhance educational and psychological assessment from various perspectives. Regarding educational testing, not only is research presented on the assessment *of* learning, that is, the summative assessment of learning outcomes (Molnár and Csapó), but a number of studies on this topic also focus conceptually and empirically on the assessment *for* learning, that is, the formative assessment providing feedback to support the learning process (Arieli-Attali et al.; Blaauw et al.; Csapó and Molnár; den Ouden et al.; Kroeze et al.).

Table 1 gives an overview of all the papers included in this Research Topic and summarizes them with respect to their key features. Reflecting the scope of the Research Topic, we used four major categories to classify the papers: (1) papers focusing on the use of new data types and sources, (2) innovative item designs, (3) innovative test designs, and (4) statistical approaches. We refrained from multiple category assignments of papers, which was possible, and focused on their core contribution. The papers' key findings and advancements impressively represent the current state-of-the-art in the field of technology-based assessment in (standardized) educational testing, and, as topic editors, we were happy to receive such a great collection of papers with various foci.

Regarding the future of technology-based assessment, we assume that inferences about the individual's or learner's knowledge, skills, or other attributes will increasingly be based on empirical (multimodal) data from less- or non-standardized testing situations. Typical examples are stealth assessments in digital games (Shute and Ventura, 2013; Shute, 2015), digital learning environments (Nguyen et al., 2018), or online activities (Kosinski et al., 2013). Such new kinds of unobtrusive, continuous assessments will further extend the traditional assessment paradigm and enhance our understanding of what an item, a test, and the empirical evidence for inferring attributes can be (Mislevy, 2019). Major challenges lie in the identification and synthesis of evidence from the situations the individual encounters in these non-standardized settings, as well as in validating the interpretation of derived measures. This Research Topic provides much input for these questions. We hope that you will enjoy reading the contributions as much as we did.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

This work was funded by the Centre for International Student Assessment (ZIB) in Germany. We thank all authors who have contributed to this Research Topic and the reviewers for their valuable feedback on the manuscript.

REFERENCES

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl. Acad. Sci. U.S.A.* 110, 5802–5805. doi: 10.1073/pnas.1218772110

Mislevy, R. (2019). "On integrating psychometrics and learning analytics in complex assessments," in *Data Analytics and Psychometrics,* eds H. Jiao, R. W. Lissitz, and A. van Wie (Charlotte, NC: USA Information Age Publishing), 1–52.

Nguyen, Q., Huptych, M., and Rienties, B. (2018). "Linking students' timing of engagement to learning design and academic performance," in *Paper presented at the Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (Sydney, NSW).

Shute, V. J. (2015). "Stealth assessment," in *The SAGE Encyclopedia of Educational Technology*, ed J. Spector (Thousand Oaks, CA: SAGE Publications, Inc.), 675–676.

Shute, V. J., and Ventura, M. (2013). *Stealth Assessment: Measuring and Supporting Learning in Video Games*. Cambridge, MA: MIT Press.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Psychology

TECHNOLOGY REPORT
published: 22 November 2018
doi: 10.3389/fpsyg.2018.02241

Assessment of Unilateral Spatial Neglect Using a Free Mobile Application for Italian Clinicians

Pietro Cipresso[1,2†], Elisa Pedroli[1†], Silvia Serino[1,2], Michelle Semonella[1], Cosimo Tuena[1], Desirée Colombo[3], Federica Pallavicini[4] and Giuseppe Riva[1,2]*

[1] Applied Technology for Neuro-Psychology Lab, Istituto Auxologico Italiano, Milan, Italy, [2] Department of Psychology, Università Cattolica del Sacro Cuore, Milan, Italy, [3] Department of Basic Psychology, Clinic and Psychobiology, Universitat Jaume I, Castellón de la Plana, Spain, [4] "Riccardo Massa" Department of Human Sciences for Education, University of Milano-Bicocca, Milan, Italy

OPEN ACCESS

Edited by:
Ronny Scherer,
University of Oslo, Norway

Reviewed by:
Timothy R. Brick,
Pennsylvania State University, United States
Nelson Silva Filho,
Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP), Brazil

***Correspondence:**
Elisa Pedroli
e.pedroli@auxologico.it

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

***Received:** 30 April 2018*
***Accepted:** 29 October 2018*
***Published:** 22 November 2018*

Citation:
Cipresso P, Pedroli E, Serino S, Semonella M, Tuena C, Colombo D, Pallavicini F and Riva G (2018) Assessment of Unilateral Spatial Neglect Using a Free Mobile Application for Italian Clinicians. Front. Psychol. 9:2241. doi: 10.3389/fpsyg.2018.02241

Background: Unilateral Spatial Neglect (USN) is traditionally assessed with paper-and-pencil tests or computer-based tests. Thanks to the wide-spreading of mobile devices, and the extensive capabilities that they have in dealing complex elements, it is possible to provide clinicians with tools for cognitive assessment. Contemporary 3D engine is, in general generally, able to deploy complex 3D environments for iOS, Android and Windows mobile, i.e., most of the mobile phone and tablet operative systems.

Results: This brand-new scenario and pressing requests from professionals, pushed us to build an application for the assessment of USN. Our first attempt was to replicate the classic cognitive tests, traditionally used at this purpose. Ecological assessment is difficult in real scenarios so we implemented virtual environments to assess patients' abilities in realistic situations. At the moment, the application is available only for iPad and iPhone for free, from the Apple Store, under the name of "Neglect App." The App contains traditional tests (e.g., barrage with and without distractors) and ecological tests (e.g., to distribute the tea in a table to close people). Scoring of each test is available to the clinicians through a database with the executed ecological tasks, that are stored locally.

Conclusion: In conclusion, Neglect App is an advanced mobile platform for the assessment of Neglect.

Keywords: neglect, psychometrics, computational psychometrics, ecological assessment, mobile virtual reality, mHealth, pervasive computing, mental health

INTRODUCTION

The Unilateral Spatial Neglect (USN) or Neglect manifests in about 2/3 of patients during the acute phase following a stroke. Stroke is an occurrence of cerebral vascular disease resulting in acute disruption of the focal or generalized brain function. Every year, there are approximately 500,000 stroke patients in Europe. This is the third leading cause of death in Western countries after cardiovascular diseases and malignancies (Sudlow and Warlow, 1997; Pendlebury et al., 2009; Roger et al., 2012; Go et al., 2014; Mozaffarian et al., 2015).

A stroke is a catastrophic and often unexpected event with a wide range of physical and psychological consequences in the long term for both patients and their families.

The long-term effects of stroke depend on the type, severity, and location of the occlusion: it is important to identify as soon as possible which part of the brain and how severely it has been affected. In general, two basic categories of impairments or disabilities can be identified: cognitive disability, which includes memory problems, difficulty in executive functions and aphasia, and motor disabilities, which includes the inability to walk and problems with coordination and balance (ataxia), mobility difficulties with arms, hemiparesis or hemiplegia, spasticity and contractures.

In particular USN can be defined as a disorder because the patient has difficulties to explore, pay attention, perceive, and act within the space opposite the region of the brain lesion. Often, there is also a difficulty in elaborating mental images in the opposite side of the damaged one. It is important to underline that the problems shown by patients are not caused by primary sensory or motor deficit, although they are often associated with hemiplegia and hemianopia (Ducros, 2012; Vocat et al., 2013; Heilman, 2014; Saj et al., 2014). These problems occur mainly following a damage to the right brain hemisphere, but there are patients in which the syndrome arose after a left-sided lesion; right neglect is considered less severe and less enduring (Stone et al., 1991; Halligan and Robertson, 2014). Regardless the side of the lesion, this disorder can be caused by the damage of several areas; the most typical one is the parietal lobe, specifically the inferior parietal lobule, followed by the frontal lobe and other sub-cortical structures such as the thalamus and the basal ganglia (Moretti et al., 2012; Saj et al., 2012; Antal et al., 2014).

In the acute phase and in the more severe form, the patient appears with the head and the gaze turned to the ipsilesional side, insensitive to any stimulation coming from the contralesional side. Over time, symptoms may ease, although more and more studies are showing that the disorder can last even for years (Kerkhoff and Schenk, 2012).

The neglect can be accompanied by several phenomena:

- Anosognosia: the unawareness of his/her own disability does not allow the patient to formulate strategies to compensate the lack of exploration in the left space (something that occurs when patients are affected by hemiplegia or hemianopsia; Bisiach et al., 1986; Pia et al., 2004).
- Anosodiaphoria: indifference or inadequate emotional response showing as the awareness of the disease increases (Barré et al., 1923; Migliaccio et al., 2014; Gasquoine, 2015).
- Extinction to double stimulation: the patient, who is able to identify a single stimulus presented on the contralesional side, cannot recognize it if presented together with an ipsilesional one (Vossel et al., 2011; de Haan et al., 2012; Heidler-Gary et al., 2013).
- Allochiria: the patient transports a stimulus to the neglected side to the ipsilesional one. For example, in case of left-sided neglect, if touched on his/her left leg the patient mentions to have being touched on the right one (Treccani et al., 2012; Antoniello and Gottesman, 2013; Marshall et al., 2013; Bartolomeo, 2014).

The standard neuropsychological tests for the analysis of extra personal neglect can be divided into:

- *cancellation test*: tasks requiring that the patient deletes certain elements within a spreadsheet, alone or mixed with distracters (Albert, 1973; Diller et al., 1974; Halligan and Marshall, 1989).
- *reading test*: both words and phrases (Pizzamiglio et al., 1989a,b) evaluate what is called neglect dyslexia.
- *bisection of lines*: the patient is required to mark the half of lines of different lengths and place in different ways in the space.
- *copy of drawings*: the patient has to copy a complex figure such as a daisy.

What may be of great help to clinical placement is real exploration of space such as the room, where the patient is hospitalized to or the one where tests are conducted to have more complete picture of the patient's spatial abilities. Unfortunately, it is difficult to make these tests in a clinical setting because of the higher requested time and human resources. Finally, it is difficult to standardize these tests due to the heterogeneity of the experimental situations.

The purpose of this App was to include the described tests for a portable and electronic use, including also an automated score recording, that can also help in simplifying the difficult process of neuropsychological assessment. In one hand several tests have been included, as it is shown in the following sections. On the other hand, we made the effort of including new paradigms and tests that are difficult to be made in paper and pencil mode. In particular, navigation tasks and ecological tests represent our effort integrating current paradigms for neuropsychological assessment. At the moment, a plaint of possible features and indexes are probably still missing, however, the App represent the first effort ever in integrating many tests and tasks in a mobile application. This could be the first step toward future integrations.

IMPLEMENTATION

Neglect App is the first application for mobile devices which makes use of the huge potential of virtual environments for the assessment of the USN for which an evaluation as effective and prompt as possible are crucial (Pallavicini et al., 2015; Pedroli et al., 2015a,b, 2016).

During the process of the design of the App, we also exploited the potential of 3D interactive applications for preventing and/or improving cognitive impairments related to USN, on the basis of a series of advantages amply documented by scientific literature:

Neuroplasticity

Neuroplasticity: the App permits to use scenarios specifically designed following principles that regulate and facilitate neuroplasticity (the neurobiological process basis of recovery of cognitive and motor functions), such as exercise intensity, exercise frequency, "enriched stimulation" (Cheung et al., 2014; Ekman et al., 2018).

Personalized Training

Personalized training: the App is based on highly automated functioning mechanisms that requires a minimal contribution by the clinical therapists, who have the possibility to customize the intensity and the difficulty of the training based on the specific needs of the patients; Engaging tasks: in the App, the content of training exercises are based on defining some tasks to re-train specific abilities (for example, increasing complexity time by time), and in the same time integrating in the scenario some recreational elements to maintain a high level of engagement and compliance of the older participant. Specifically, ecological simulations can be particularly engaging by supporting a process known as "transformation of flow," defined as a person's ability to exploit an optimal (flow) experience to identify and use new and unexpected psychological resources as sources of involvement (Riva et al., 2006; Pedroli et al., 2018). Also, presence is a key point of the engagement in the use of technology. Presence is usually defined as the "sense of being there" or the "feeling of being in a world that exists outside the self." The ability to interact actively with the environment greatly improves the possibility of experiencing presence (Riva et al., 2007; Villani et al., 2012b).

Tracking and Objective/Quantitative Measure

Tracking and objective/quantitative measure: it is possible to record a high quantity of data and use them to create some indexes of performance in order to measure in a quantitative and objective way the improvement of the performances observable in the course of possible rehabilitative process.

Transferring of the Training in Activity of Daily Living (ADL)

Transferring of the training in activity of daily living (ADL): many studies suggested the potential offered by ecological tasks to transfer the results of re-learning of cognitive and motor abilities that were damaged in ADL. Positive impact of ecological tasks on ADL is documented by many studies (Laver et al., 2015; Chiang et al., 2017).

A previous pilot study investigating the correlation between Neglect App test and classic test in order to understand the usability and ecologicity of our app. Results showed that the cancelation tests of Neglect App were equally effective to the traditional tests in the screening of symptoms between patients with and without neglect. Moreover, the Neglect App Card Dealing task was more sensitive in detecting neglect symptoms than traditional functional task (Pallavicini et al., 2015).

Neglect App contains a series of trials for neglect evaluation through classic tests and virtually interactive environments with the double advantage of automating and making more ecological the evaluation of neglect patients, who thus show a difficulty and/or incapacity to explore, pay attention, perceive, and act in the space region opposite to the area of the brain lesion. Thanks to Neglect App, it is possible to evaluate the explorative behavior of the patient in a fast and simple way, inside ecological environments and receiving all the data, from the performed sessions, included in a database. Neglect App can be downloaded for free at: https://itunes.apple.com/it/app/neglect-app/id788480837?mt=8.

RESULTS AND DISCUSSION

Evaluation is composed of nine exercises divided in two groups: ecological tasks and barrage. The first group comprises ecological tasks, some of which inspired by the ecological battery by Zoccolotti et al., 1994), some others created by starting from real life situations and tasks used clinically but lacking standardization. The app always provides all the score dividing the results in left, right, center, and total areas. Moreover a screenshot with the results is always recorded and generated in the report.

Ecological Tasks

Serve Tea

The patient is required to distribute tea to himself and people sitting at the table with him using objects placed in the center of the table (**Figure 1**).

The task is commonly used by clinicians in real settings, however, the experience has been replicated in the tablet to be more usable, keeping its own ecological validity.

The patients can be used their finger to drag and drop the single objects for taking the task as requested by the App.

The App already contains the instructions that have to be followed, so the clinicians have just to give the tablet to the patient observing the correct use while executing the task.

Clinicians are not required to take note of the performed actions since the App is able to record every significant action consequently calculating the standard scores that can be used and integrated in a clinical protocol.

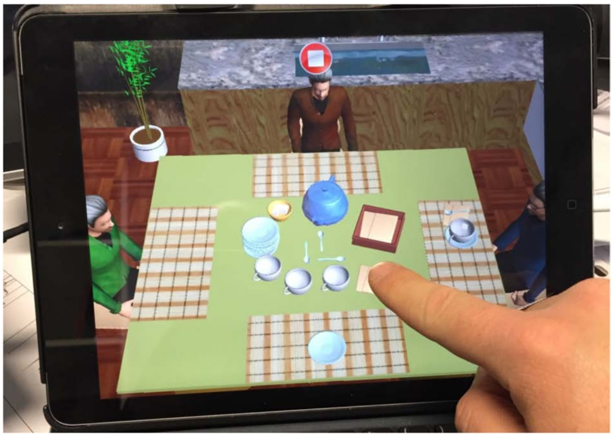


FIGURE 1 | Laying the table tapping on the iPad using the Neglect App.

The score is assigned on the basis of proper and wrong objects placed to the right, in the center and to the left. Time employed and unconsidered objects are also signaled (**Figure 2**).

In any time, clinicians are able to access to the patients' score directly from the App, visualizing each score in each task assigned at any time.

Card Dealing

The second exercise requires the patient to hand out playing cards to himself and people sitting at the table with him (**Figure 3**). The score is assigned on the basis of correctly given cards, omitted cards, wrong cards (i.e., those in excess) to the right, to the left and in the middle and the time employed to compete the exercise.

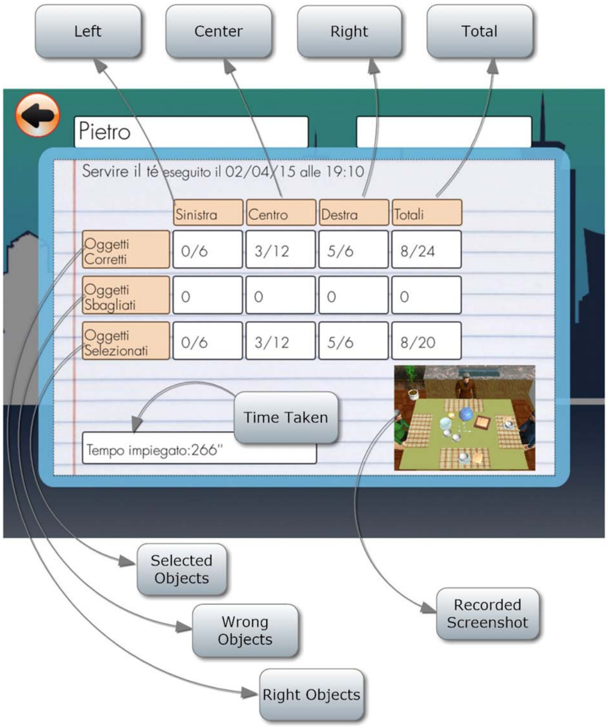
Controlling an Orders List

In this task, the patient is required to check an orders list to verify if the dishes noted herein are on the shelves; if they are, he/she will have to select the dish on the shelf and the note on the list (**Figure 4**).

Score is assigned on the basis of: the dishes selected correctly; those selected wrongly; the correct dishes omitted; the correct selections and omissions on the list; and the time taken (**Figure 5**).

Exploration

Within this environment the patient finds him/herself in a room in which he/she can move freely to left or right describing all the objects that are in the room and touching them accordingly (**Figure 6**). The app calculates automatically, as the patient moves, if the selected object was on the right or the left. The report indicates selected objects on the left, the ones on the right, repetitions, time employed, and omitted elements.



Pietro

Servire il té eseguito il 02/04/15 alle 19:10

	Sinistra	Centro	Destra	Totali
Oggetti Corretti	0/6	3/12	5/6	8/24
Oggetti Sbagliati	0	0	0	0
Oggetti Selezionati	0/6	3/12	5/6	8/20

Tempo impiegato:266''

FIGURE 2 | Scores report for the exercise "Distribute the tea."

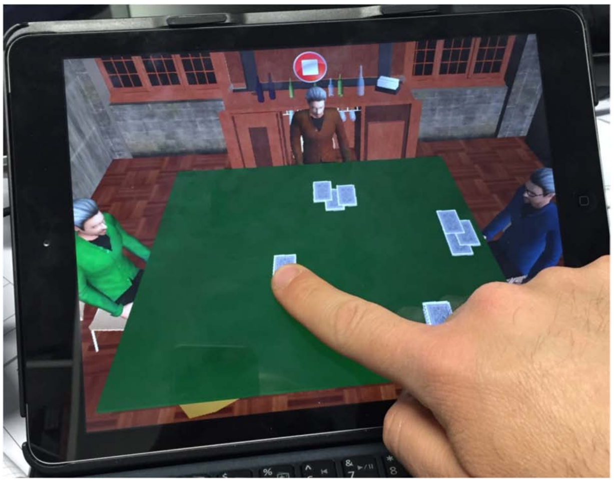
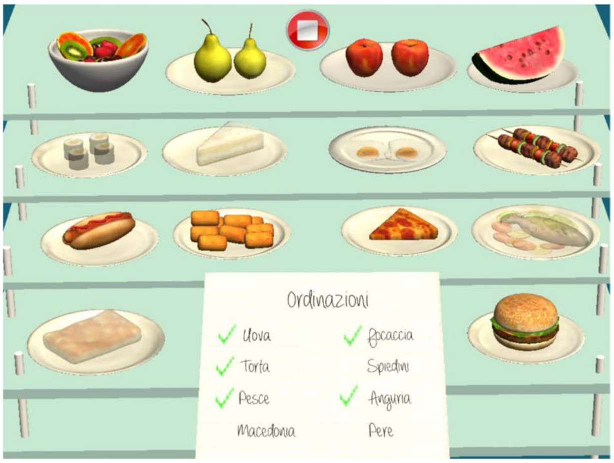


FIGURE 3 | Distributing cards tapping on the iPad using the Neglect App.



Ordinazioni

✓ Uova ✓ Focaccia
✓ Torta Spiedini
✓ Pesce ✓ Anguria
Macedonia Pere

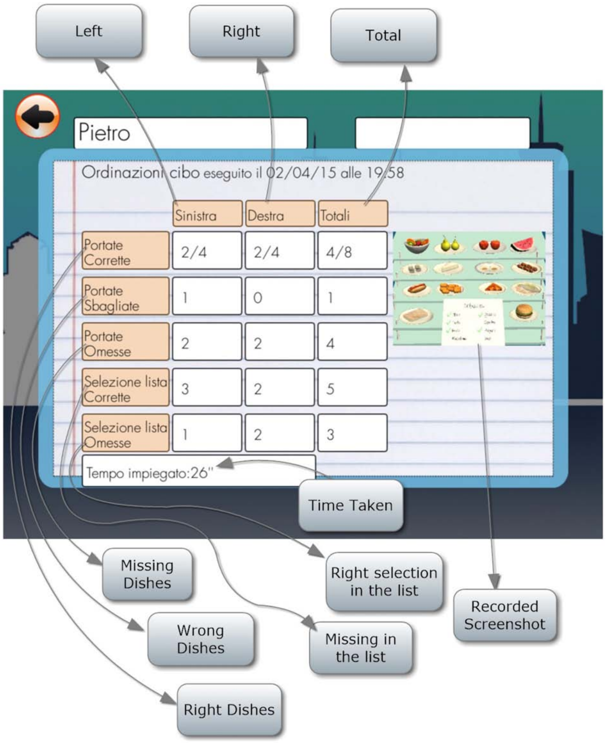
FIGURE 4 | Controlling an orders list task.

Apples Pursuit

Within this environment the patient finds him/herself in an office in which he/she can move freely to left or right to identifying and touching all the apples inside (**Figure 7**). The app calculates automatically, as the patient moves, if the selected apple was on the right or the left. The report indicates selected apples on the left, the ones on the right, repetitions, time employed, and omitted apples.

Barrage Tasks

Barrage tests take the cue from classical cancelation tasks commonly used clinically (Zoccolotti et al., 1994) and comprehend four exercises, described below.



Pietro

Ordinazioni cibo eseguito il 02/04/15 alle 19:58

	Sinistra	Destra	Totali
Portate Corrette	2/4	2/4	4/8
Portate Sbagliate	1	0	1
Portate Omesse	2	2	4
Selezione lista Corrette	3	2	5
Selezione lista Omesse	1	2	3

Tempo impiegato:26''

FIGURE 5 | Controlling an orders list score.

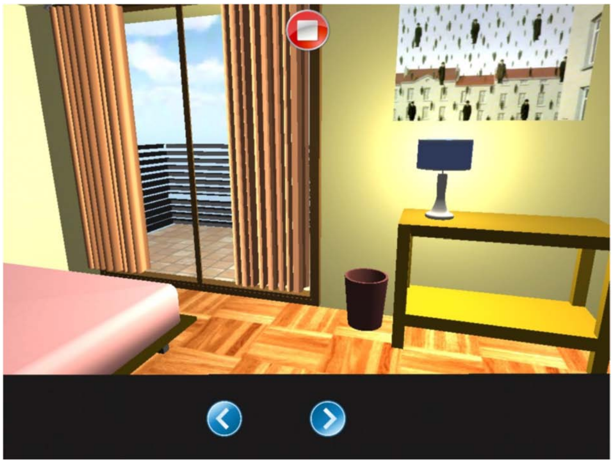


FIGURE 6 | Exploration task.

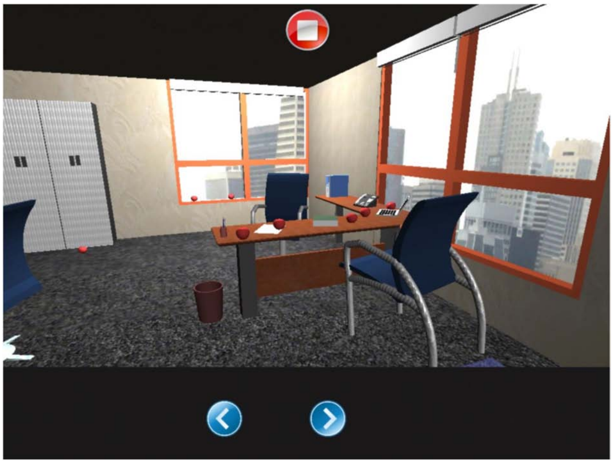


FIGURE 7 | Apples pursuit task.

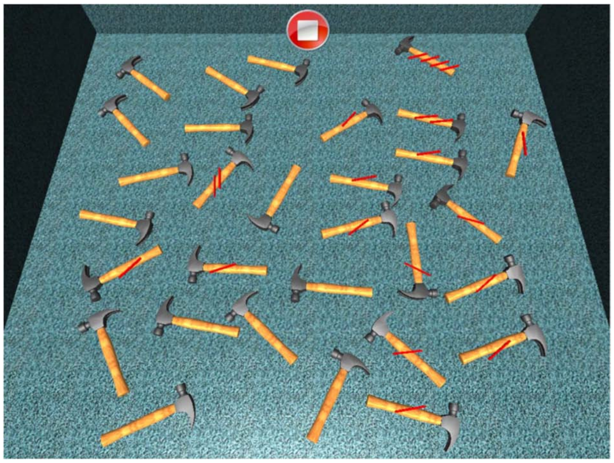


FIGURE 8 | Simple barrage task.

Simple Barrage

Patient is required to select all objects (hammers) in the room. There are no distractors (**Figure 8**). The number of selected objects, repetitions, objects omitted on the left and on the right and time employed are considered (**Figure 9**).

Simple Barrage With Distractors

Patient is required to select all target objects (screwdrivers) in the room, which are mixed with distractors (**Figure 10**). The number of selected target objects, repetitions, target objects omitted and the distractors selected on the left and on the right and time employed are considered (**Figure 11**).

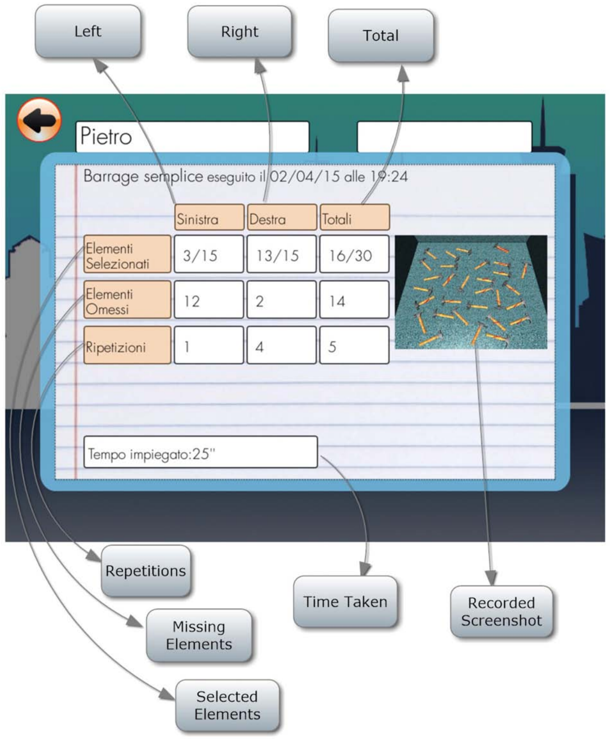
Dynamic Barrage

Patient is required to select all objects (balloons) in the sky. There are no distractors. The peculiarity here is that the objects are moving (**Figure 12**). The number of selected objects, repetitions, objects omitted on the left and on the right and time employed are considered (**Figure 13**).

Dynamic Barrage With Distractors

Patient is required to select all target objects (kites) in the room, which are mixed with distractors (**Figure 14**). The number of selected target objects, repetitions, target objects omitted and the distractors selected on the left and on the right and time employed are considered (**Figure 15**).

A qualitative analysis of the barrage tasks may give information about dysexecutive behaviors because it is possible



Pietro

Barrage semplice eseguito il 02/04/15 alle 19:24

	Sinistra	Destra	Totali
Elementi Selezionati	3/15	13/15	16/30
Elementi Omessi	12	2	14
Ripetizioni	1	4	5

Tempo impiegato:25''

FIGURE 9 | Simple barrage score.

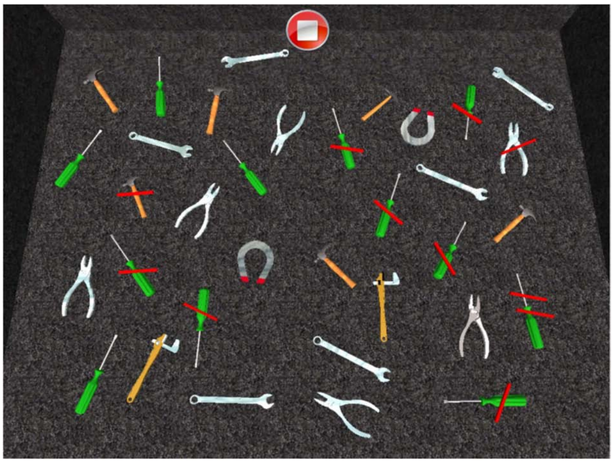
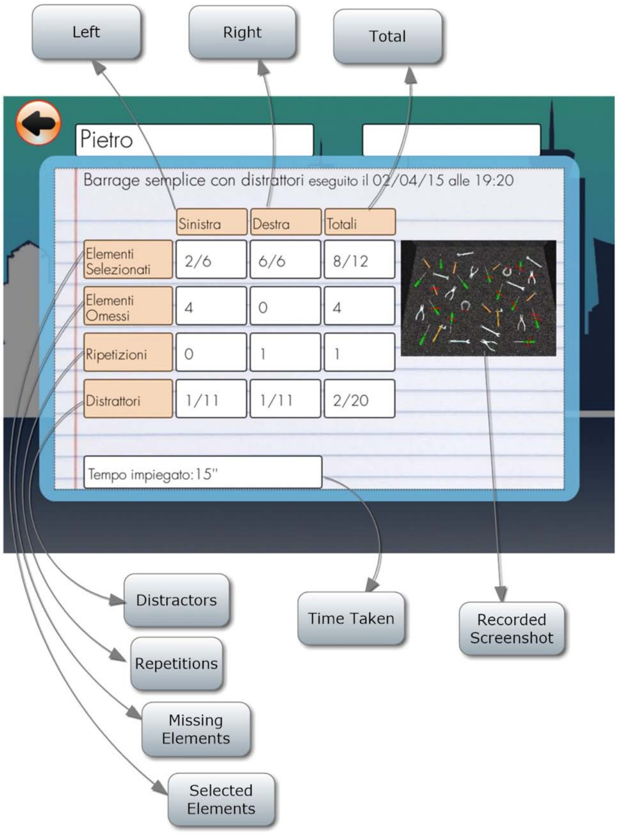


FIGURE 10 | Simple barrage with distractors task.

to select multiple times every single item and the target in the simple version of both barrage tasks (simple and dynamic) is in the environment of the barrage with distractions tasks.



Pietro

Barrage semplice con distrattori eseguito il 02/04/15 alle 19:20

	Sinistra	Destra	Totali
Elementi Selezionati	2/6	6/6	8/12
Elementi Omessi	4	0	4
Ripetizioni	0	1	1
Distrattori	1/11	1/11	2/20

Tempo impiegato:15''

FIGURE 11 | Simple barrage with distractors score.

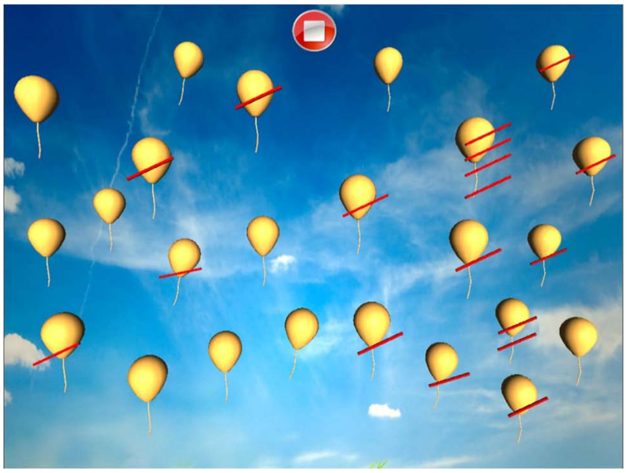
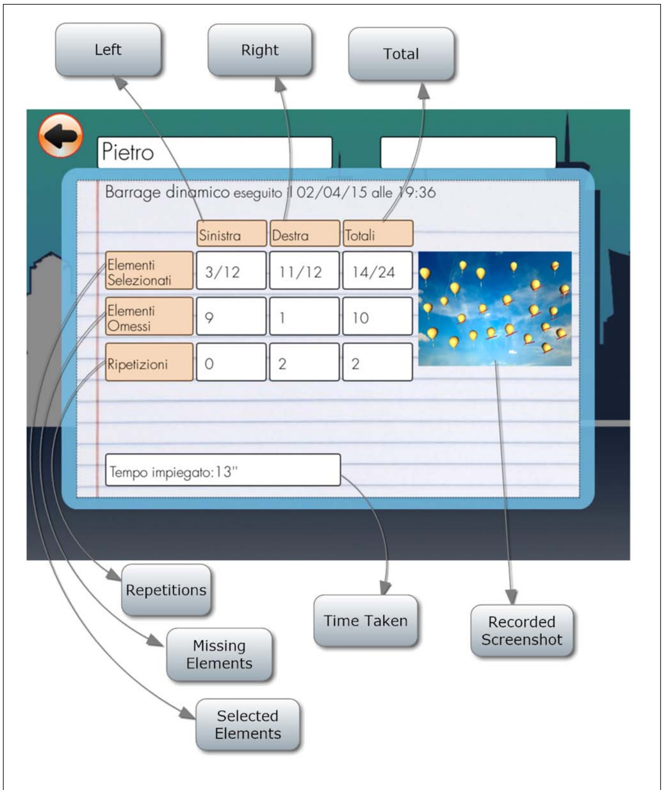


FIGURE 12 | Dynamic barrage task (the balloons are in a continuous movement).

Data Management

All data can be downloaded in a unique file by connecting the iPad to a Computer or a Mac equipped with iTunes software. Once downloaded, the file can be easily read with a client software able to interact with SQL Databases (**Figure 16**). All data, including images, are exportable to be computed for the statistical analysis.



Pietro

Barrage dinamico eseguito il 02/04/15 alle 19:36

	Sinistra	Destra	Totali
Elementi Selezionati	3/12	11/12	14/24
Elementi Omessi	9	1	10
Ripetizioni	0	2	2

Tempo impiegato:13''

FIGURE 13 | Dynamic barrage score.

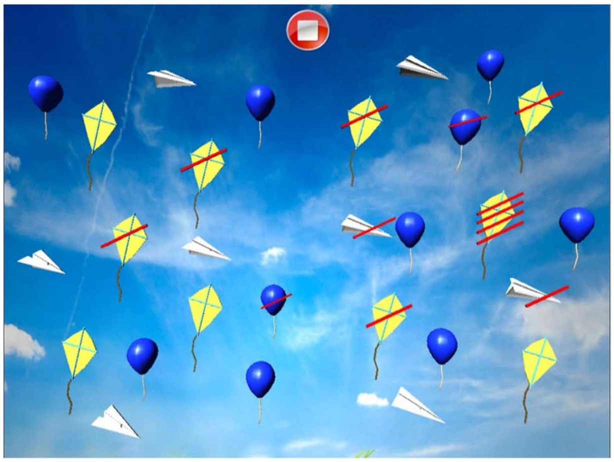
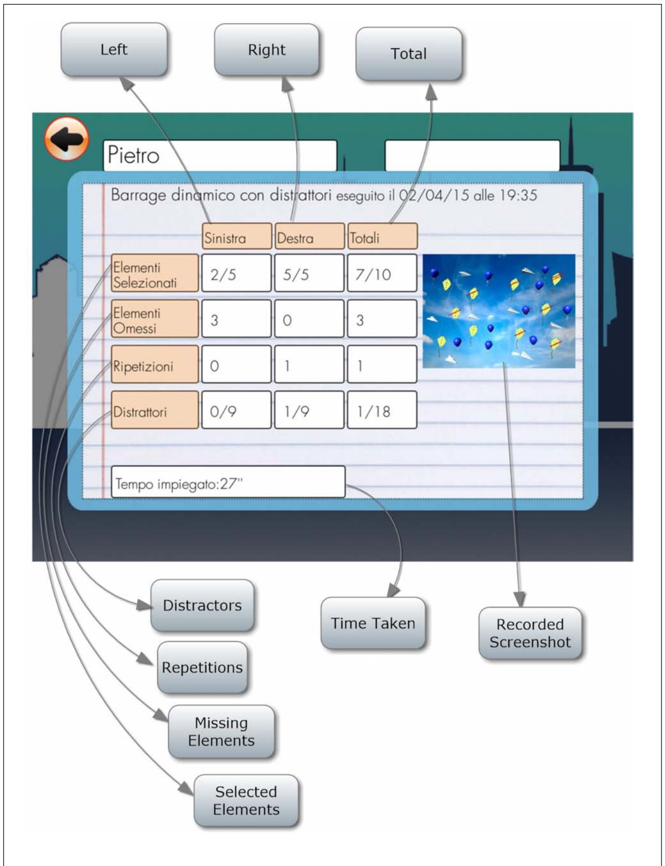


FIGURE 14 | Dynamic barrage with distractors task (all the elements are in a continuous movement).

CONCLUSION

Neglect may influence the behaviors of the patient in everyday life activity: they can constantly hit the objects placed on his left, not paying attention to the left side of the road when he crosses. In severe cases he can ignore the food in the left half of the plate. So, it has a sufficiently serious framework that allows the patient to cope independently.



Pietro

Barrage dinamico con distrattori eseguito il 02/04/15 alle 19:35

	Sinistra	Destra	Totali
Elementi Selezionati	2/5	5/5	7/10
Elementi Omessi	3	0	3
Ripetizioni	0	1	1
Distrattori	0/9	1/9	1/18

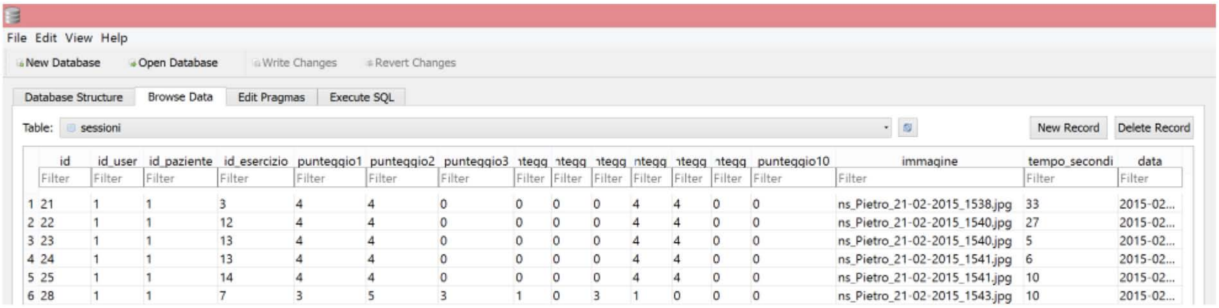
Tempo impiegato:27''

FIGURE 15 | Dynamic barrage with distractors score.

The functions such as memory, speech, or attention in neuropsychological research were traditionally assessed through program of standardized tests, which have clear psychometric advantages, but often measure behaviors that are very different from those of everyday life (Chaytor and Schmitter-Edgecombe, 2003).

In recent years, there has been a growing interest in the development of tools that allows ecological and functional assessment above all by using mobile device (Villani et al., 2012a, 2013; Carbonaro et al., 2014; Pedroli et al., 2015b). The results of a meta-analytic review of Neguț et al. (2016) support the sensitivity of virtual reality tools in detecting cognitive deficit. One of the areas where emerges this need is the assessment of neglect. We decided to diffuse the application in the Italian market with a future intention to extend worldwide a possible English version. The Neglect App temporal cycle concern from the moment of the patients into the Clinique to the continuous assessment at the patient's home and back to the Clinique in a closed loop for the continuous assessment (**Figure 17**).

Assessment by using a mobile tool and virtual environments might represent a great challenge for very sophisticated methods able to assess in a way before unthinkable and sometimes impossible in real settings. In particular, navigation tasks allow the system to identifying if an object in the space is located in left



	id	id_user	id_paziente	id_esercizio	punteggio1	punteggio2	punteggio3	ıtegg	ıtegg	ıtegg	ntegg	ıtegg	ıtegg	punteggio10	immagine	tempo_secondi	data
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	21	1	1	3	4	4	0	0	0	0	4	4	0	0	ns_Pietro_21-02-2015_1538.jpg	33	2015-02...
2	22	1	1	12	4	4	0	0	0	0	4	4	0	0	ns_Pietro_21-02-2015_1540.jpg	27	2015-02...
3	23	1	1	13	4	4	0	0	0	0	4	4	0	0	ns_Pietro_21-02-2015_1540.jpg	5	2015-02...
4	24	1	1	13	4	4	0	0	0	0	4	4	0	0	ns_Pietro_21-02-2015_1541.jpg	6	2015-02...
5	25	1	1	14	4	4	0	0	0	0	4	4	0	0	ns_Pietro_21-02-2015_1541.jpg	10	2015-02...
6	28	1	1	7	3	5	3	1	0	3	1	0	0	0	ns_Pietro_21-02-2015_1543.jpg	10	2015-02...

FIGURE 16 | Database management to brows and analyze the data collected.

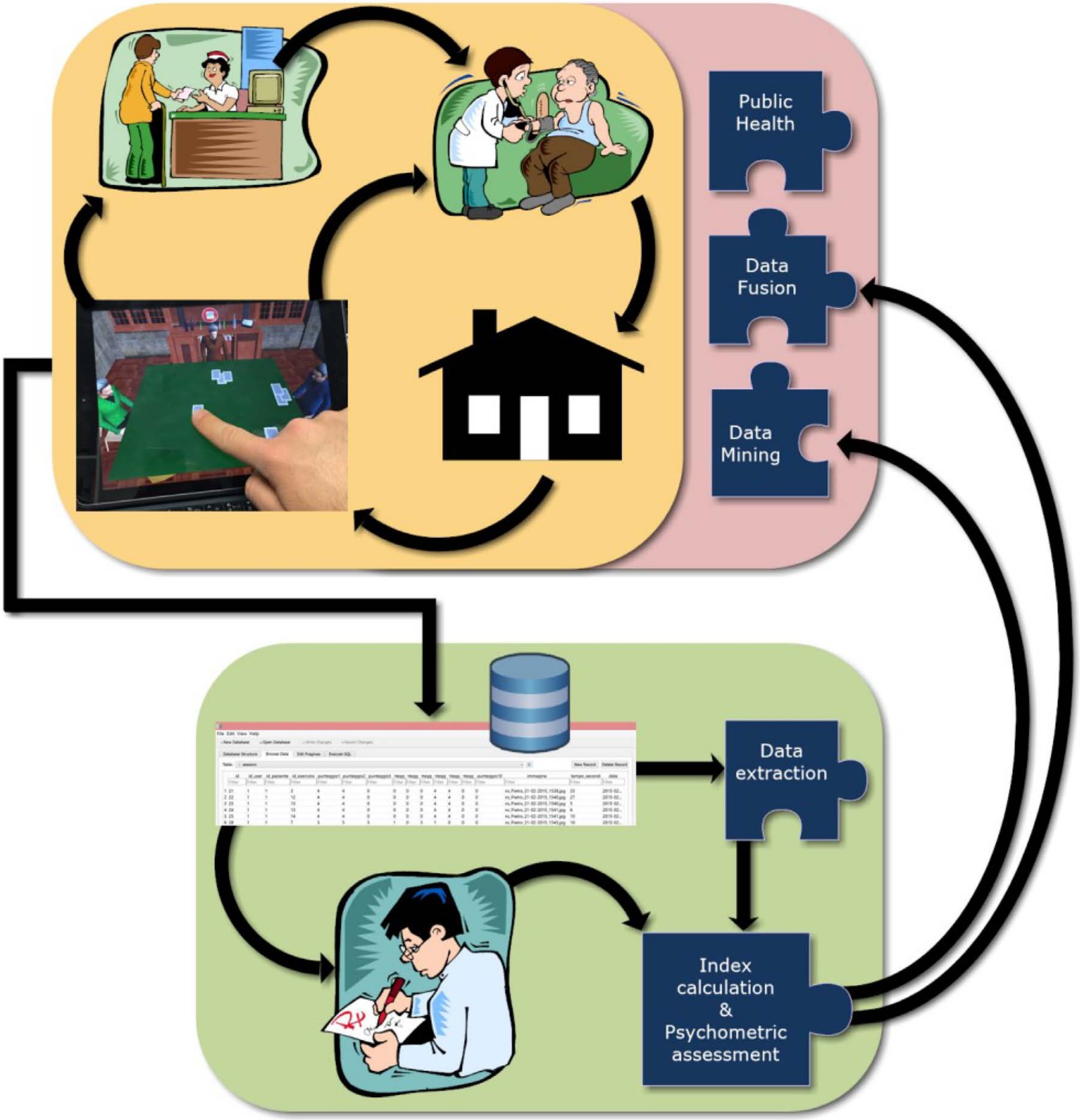


FIGURE 17 | The Neglect App cycle. From the patients into the clinic to the continuous assessment at the patient's home.

or right side when selected. On the other hand, in real settings to do this navigation task is too expensive, requiring eye-tracking glasses. Moreover a computational approach can be easily used to provide more feedback to the patients and to model behaviors (Cipresso, 2015; Cipresso et al., 2015).

One of the limitations of the App is the screen dimension, that does not provide any direct advantage compared to paper and pencil test. Actually, this limitation has been recently overcome by the iPad Pro 12,9″ that can be effectively used with our App, being totally compatible. Another limitation is the lack of normative data available for a quantitative analysis of the results; only a qualitative analysis is recommended. A future study could be able to fill this gap.

At the moment we have not implement some additional information and indexes that could help the clinicians to better understand the characteristics of their explorative behaviors in order to program a more personalized rehabilitations. In particular, it could be interesting to report the starting point and the path of the exploration made by patients or some other indexes like the ones reported in the Chung et al.'s (2016) article.

Additionally, to create some tasks for rehabilitations could make our application completer and more interesting. Provide some tasks for make exercises in a virtual environment could help patients and clinicians to improve clinical practice.

The future development will have directed to fill these limitations with the addition of some specific tasks both for assessment and rehabilitation. A manipulation of the cognitive complexity of the barrage tasks according to the criteria proposed by Ricci et al. (2016) and Sarri et al. (2009) could help to have a more precise assessment process. To aim this scope the introduction of a 3D version of line bisection task are also consider because some patients may show neglect symptoms in this kind of task and not in the barrage one.

Also, a new version developed to take advantage of the immersive technology could be designed in order to reach a higher degree of ecologicity.

After all these modifications a validation study will be necessary in order to prove the validity of our system. Also, a clinical trial for the rehabilitation session could be done in order to prove the usefulness of a computerizing protocol. Both, convergent and discriminant validity, need to be verified comparing current tools accordingly. At this purpose can be used current neuropsychological battery and specific test, such as barrage test, front assessment battery, real task (e.g., lay the table in real context), and so on.

We are so providing the scientific and clinical communities a free advanced tool able to be a practical and flexible way for the assessment directly in the patients' place but also a brand-new way for the assessment of Neglect.

Availability and Requirements

- **Project name:** Neglect App.
- **Project home page:** https://itunes.apple.com/it/app/neglect-app/id788480837?mt=8.
- **Operating system(s):** iOS Platform (at least iOS 6.0 is required).
- **Programming language:** No programming language is required for using the App. The Neglect App has been developed by using Unity.
- **Other requirements:** the App works also on iPhone device but iPad device is suggested for the best use and visualization.
- **License:** Available for free.
- **Any restrictions to use by non-academics:** No restrictions.

AUTHOR CONTRIBUTIONS

PC and EP wrote the manuscript. PC, EP, SS, MS, CT, and DC collected the literature materials. GR supervised the study. PC, EP, SS, FP, and GR conceived the idea of the study, established the software requirements, and supervised the technological, clinical, and scientific aspects. All authors read and approved the final manuscript.

FUNDING

This study was partially supported by the Italian funded project "VRehab. Virtual Reality in the Assessment and TeleRehabilitation of Parkinson's Disease and Post-Stroke Disabilities" – RF-2009-1472190.

REFERENCES

Albert, M. L. (1973). A simple test of visual neglect. *Neurology* 23, 658–664. doi: 10.1212/WNL.23.6.658

Antal, M., Beneduce, B. M., and Regehr, W. G. (2014). The substantia nigra conveys target-dependent excitatory and inhibitory outputs from the basal ganglia to the thalamus. *J. Neurosci.* 34, 8032–8042. doi: 10.1523/JNEUROSCI.0236-14.2014

Antoniello, D., and Gottesman, R. (2013). Allochiria in acute right hemispheric dysfunction (P01. 009). *Neurology* 80.

Barré, J., Morin, L., and Kaiser, J. (1923). Etude clinique d'un nouveau cas d'anosognosie de Babinski. *Rev. Neurol.* 39, 500–503.

Bartolomeo, P. (ed.) (2014). "Unilateral spatial neglect: clinical aspects," in *Attention Disorders After Right Brain Damage*, (London: Springer), 49–83. doi: 10.1007/978-1-4471-5649-9_4

Bisiach, E., Vallar, G., Perani, D., Papagno, C., and Berti, A. (1986). Unawareness of disease following lesions of the right hemisphere: anosognosia for hemiplegia and anosognosia for hemianopia. *Neuropsychologia* 24, 471–482.

Carbonaro, N., Cipresso, P., Tognetti, A., Anania, G., De Rossi, D., Pallavicini, F., et al. (2014). Psychometric assessment of cardio-respiratory activity using a mobile platform. *Int. J. Handheld Comput. Res.* 5, 13–29. doi: 10.4018/ijhcr.2014010102

Chaytor, N., and Schmitter-Edgecombe, M. (2003). The ecological validity of neuropsychological tests: a review of the literature on everyday cognitive skills. *Neuropsychol. Rev.* 13, 181–197.

Cheung, K. L., Tunik, E., Adamovich, S. V., and Boyd, L. A. (2014). "Neuroplasticity and virtual reality," in *Virtual Reality for Physical and Motor Rehabilitation*, eds P. L. Weiss, E. A. Keshner, and M. F. Levin (New York, NY: Springer), 5–24. doi: 10.1023/B:NERV.0000009483.91468.fb

Chiang, V. C.-L., Lo, K.-H., and Choi, K.-S. (2017). Rehabilitation of activities of daily living in virtual environments with intuitive user interface and force feedback. *Disabil. Rehabil.* 12, 672–680. doi: 10.1080/17483107.2016.1218554

Chung, S. J., Park, E., Ye, B. S., Lee, H. S., Chang, H.-J., Song, D., et al. (2016). The computerized table setting test for detecting unilateral neglect. *PLoS One* 11:e0147030. doi: 10.1371/journal.pone.0147030

Cipresso, P. (2015). Modeling behavior dynamics using computational psychometrics within virtual worlds. *Front. Psychol.* 6:1725. doi: 10.3389/fpsyg.2015.01725

Cipresso, P., Matic, A., Giakoumis, D., and Ostrovsky, Y. (2015). Advances in computational psychometrics. *Comput. Math. Methods Med.* 2015:418683. doi: 10.1155/2015/418683

de Haan, B., Karnath, H.-O., and Driver, J. (2012). Mechanisms and anatomy of unilateral extinction after brain injury. *Neuropsychologia* 50, 1045–1053. doi: 10.1016/j.neuropsychologia.2012.02.015

Diller, L., Ben-Yishay, Y., Gerstman, L. J., Goodkin, R., Gordon, W., and Weinberg, J. (1974). *Studies in Cognition and Rehabilitation in Hemiplegia: Rehabilitation Monograph*. New York, NY: University Medical Center.

Ducros, A. (2012). Reversible cerebral vasoconstriction syndrome. *Lancet Neurol.* 11, 906–917. doi: 10.1016/S1474-4422(12)70135-7

Ekman, U., Fordell, H., Eriksson, J., Lenfeldt, N., Wåhlin, A., Eklund, A., et al. (2018). Increase of frontal neuronal activity in chronic neglect after training in virtual reality. *Acta Neurol. Scand.* 138, 284–292. doi: 10.1111/ane.12955

Gasquoine, P. G. (2015). Blissfully unaware: anosognosia and anosodiaphoria after acquired brain injury. *Neuropsychol. Rehabil.* 26, 261–285. doi: 10.1080/09602011.2015.1011665

Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Blaha, M. J., et al. (2014). Heart disease and stroke statistics–2014 update: a report from the American heart association. *Circulation* 129:e28. doi: 10.1161/01.cir.0000441139.02102.80

Halligan, P. W., and Marshall, J. C. (1989). 2 Techniques for the assessment of line bisection in visuo-spatial neglect - a single case-study. *J. Neurol. Neurosurg. Psychiatry* 52, 1300–1302. doi: 10.1136/jnnp.52.11.1300

Halligan, P. W., and Robertson, I. (2014). *Spatial Neglect: A Clinical Handbook for Diagnosis and Treatment*. Hove: Psychology Press. doi: 10.4324/9781315804491

Heidler-Gary, J., Pawlak, M., Herskovits, E. H., Newhart, M., Davis, C., Trupe, L. A., et al. (2013). Motor extinction in distinct reference frames: a double dissociation. *Behav. Neurol.* 26, 111–119. doi: 10.3233/BEN-2012-110254

Heilman, K. M. (2014). Possible mechanisms of anosognosia of hemiplegia. *Cortex* 61, 30–42. doi: 10.1016/j.cortex.2014.06.007

Kerkhoff, G., and Schenk, T. (2012). Rehabilitation of neglect: an update. *Neuropsychologia* 50, 1072–1079. doi: 10.1016/j.neuropsychologia.2012.01.024

Laver, K. E., George, S., Thomas, S., Deutsch, J. E., and Crotty, M. (2015). Virtual reality for stroke rehabilitation. *Cochrane Database Syst. Rev.* 11:CD008349. doi: 10.1002/14651858.CD008349.pub3

Marshall, J. C., Halligan, P. W., and Robertson, I. H. (2013). "Contemporary theories of unilateral neglect: a critical," in *Unilateral Neglect: Clinical and Experimental Studies*, eds I. H. Robertson and J. C. Marshall (Hove: Lawrence Erlbaum Associates Ltd.), 311–329. doi: 10.4324/9780203765258

Migliaccio, R., Bouhali, F., Rastelli, F., Ferrieux, S., Arbizu, C., Vincent, S., et al. (2014). Damage to the medial motor system in stroke patients with motor neglect. *Front. Hum. Neurosci.* 8:408. doi: 10.3389/fnhum.2014.00408

Moretti, D., Paternicò, D., Binetti, G., Zanetti, O., and Frisoni, G. B. (2012). EEG markers are associated to gray matter changes in thalamus and basal ganglia in subjects with mild cognitive impairment. *Neuroimage* 60, 489–496. doi: 10.1016/j.neuroimage.2011.11.086

Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., et al. (2015). Executive summary: heart disease and stroke statistics—2015 update a report from the American heart association. *Circulation* 131, 434–441. doi: 10.1161/CIR.0000000000000157

Neguţ, A., Matu, S.-A., Sava, F. A., and David, D. (2016). Virtual reality measures in neuropsychological assessment: a meta-analytic review. *Clin. Neuropsychol.* 30, 165–184. doi: 10.1080/13854046.2016.1144793

Pallavicini, F., Pedroli, E., Serino, S., Dell'Isola, A., Cipresso, P., Cisari, C., et al. (2015). Assessing unilateral spatial neglect using advanced technologies: the potentiality of mobile virtual reality. *Technol. Health Care* 23, 795–807. doi: 10.3233/THC-151039

Pedroli, E., Greci, L., Colombo, D., Serino, S., Cipresso, P., Arlati, S., et al. (2018). Characteristics, usability, and users experience of a system combining cognitive and physical therapy in a virtual environment: positive bike. *Sensors* 18:2343. doi: 10.3390/s18072343

Pedroli, E., Serino, S., Cipresso, P., Pallavicini, F., Giglioli, I. A. C., Guastafierro, E., et al. (2015a). "Neglect app. usability of a new application for assessment and rehabilitation of neglect," in *Proceedings of the 3rd 2015 Workshop on ICTs for improving patients rehabilitation research techniques*, (New York, NY: ACM), 139–143. doi: 10.1145/2838944.2838978

Pedroli, E., Serino, S., Cipresso, P., Pallavicini, F., and Riva, G. (2015b). Assessment and rehabilitation of neglect using virtual reality: a systematic review. *Front. Behav. Neurosci.* 9:226. doi: 10.3389/fnbeh.2015.00226

Pedroli, E., Serino, S., Giglioli, A. C., Pallavicini, F., Cipresso, P., and Riva, G. (2016). "The use of virtual reality tools for the assessment of executive functions and unilateral spatial neglect," in *Virtual Reality Enhanced Robotic Systems for Disability Rehabilitation*, eds F. Hu, J. Lu, and T. Zhang (Hershey, PA: Medical Information Science Reference), 115–140. doi: 10.4018/978-1-4666-9740-9.ch007

Pendlebury, S. T., Giles, M. F., and Rothwell, P. M. (2009). *Transient Ischemic Attack and Stroke: Diagnosis, Investigation and Management*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511575815

Pia, L., Neppi-Modona, M., Ricci, R., and Berti, A. (2004). The anatomy of anosognosia for hemiplegia: a meta-analysis. *Cortex* 40, 367–377. doi: 10.1016/S0010-9452(08)70131-X

Pizzamiglio, L., Cappa, S., Vallar, G., Zoccolotti, P., Bottini, G., Ciurlil, P., et al. (1989a). Visual neglect for far and near extra-personal space in humans. *Cortex* 25, 471–477.

Pizzamiglio, L., Judica, A., Razzano, C., and Zoccolotti, P. (1989b). Toward a comprehensive diagnosis of visual-spatial disorders in unilateral brain damaged patients. *Evaluacion Psicol.* 5, 199–218.

Ricci, R., Salatino, A., Garbarini, F., Ronga, I., Genero, R., Berti, A., et al. (2016). Effects of attentional and cognitive variables on unilateral spatial neglect. *Neuropsychol.* 92, 158–166. doi: 10.1016/j.neuropsychologia.2016.05.004

Riva, G., Castelnuovo, G., and Mantovani, F. (2006). Transformation of flow in rehabilitation: the role of advanced communication technologies. *Behav. Res. Methods* 38, 237–244. doi: 10.3758/BF03192775

Riva, G., Mantovani, F., Capideville, C. S., Preziosa, A., Morganti, F., Villani, D., et al. (2007). Affective interactions using virtual reality: the link between presence and emotions. *Cyber Psychol. Behav.* 10, 45–56. doi: 10.1089/cpb.2006.9993

Roger, V. L., Go, A. S., Lloyd-Jones, D. M., Benjamin, E. J., Berry, J. D., Borden, W. B., et al. (2012). Executive summary: heart disease and stroke statistics–2012 update: a report from the American heart association. *Circulation* 125:188. doi: 10.1161/CIR.0b013e3182456d46

Saj, A., Verdon, V., Vocat, R., and Vuilleumier, P. (2012). 'The anatomy underlying acute versus chronic spatial neglect'also depends on clinical tests. *Brain* 135, e207–e207. doi: 10.1093/brain/awr227

Saj, A., Vocat, R., and Vuilleumier, P. (2014). Action-monitoring impairment in anosognosia for hemiplegia. *Cortex* 61, 93–106. doi: 10.1016/j.cortex.2014.10.017

Sarri, M., Greenwood, R., Kalra, L., and Driver, J. (2009). Task-related modulation of visual neglect in cancellation tasks. *Neuropsychologia* 47, 91–103. doi: 10.1016/j.neuropsychologia.2008.08.020

Stone, S. P., Wilson, B., Wroot, A., Halligan, P. W., Lange, L. S., Marshall, J. C., et al. (1991). The assessment of visuo-spatial neglect after acute stroke. *J. Neurol. Neurosurg. Psychiatry* 54, 345–350. doi: 10.1136/jnnp.54.4.345

Sudlow, C., and Warlow, C. (1997). Comparable studies of the incidence of stroke and its pathological types results from an international collaboration. *Stroke* 28, 491–499. doi: 10.1161/01.STR.28.3.491

Treccani, B., Cubelli, R., Sellaro, R., Umiltà, C., and Della Sala, S. (2012). Dissociation between awareness and spatial coding: evidence from unilateral neglect. *J. Cogn. Neurosci.* 24, 854–867. doi: 10.1162/jocn_a_00185

Villani, D., Grassi, A., Cognetta, C., Cipresso, P., Toniolo, D., and Riva, G. (2012a). The effects of a mobile stress management protocol on nurses working with cancer patients: a preliminary controlled study. *Stud. Health Technol. Inform.* 173, 524–528.

Villani, D., Repetto, C., Cipresso, P., and Riva, G. (2012b). May I experience more presence in doing the same thing in virtual reality than in reality? An answer from a simulated job interview. *Interact. Comput.* 24, 265–272. doi: 10.1016/j.intcom.2012.04.008

Villani, D., Grassi, A., Cognetta, C., Toniolo, D., Cipresso, P., and Riva, G. (2013). Self-help stress management training through mobile phones: an experience with oncology nurses. *Psychol. Serv.* 10:315. doi: 10.1037/a0026459

Vocat, R., Saj, A., and Vuilleumier, P. (2013). The riddle of anosognosia: does unawareness of hemiplegia involve a failure to update beliefs? *Cortex* 49, 1771–1781. doi: 10.1016/j.cortex.2012.10.009

Vossel, S., Eschenbeck, P., Weiss, P., Weidner, R., Saliger, J., Karbe, H., et al. (2011). Visual extinction in relation to visuospatial neglect after right-hemispheric stroke: quantitative assessment and statistical lesion-symptom mapping. *J. Neurol. Neurosur. Psychiatry* 82, 862–868. doi: 10.1136/jnnp.2010.224261

Zoccolotti, P., Pizzamiglio, L., Pittau, P., and Galati, G. (1994). *Batteria di Test Per L'esame dell'Attenzione*. Roma: Psytest.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PERSPECTIVE
published: 12 December 2018
doi: 10.3389/fpsyg.2018.02532

Virtual Reality as a New Approach for Risk Taking Assessment

Carla de-Juan-Ripoll, José L. Soler-Domínguez, Jaime Guixeres, Manuel Contero, Noemi Álvarez Gutiérrez and Mariano Alcañiz*

Instituto de Investigación e Innovación en Bioingeniería (i3B), Universitat Politècnica de València, Valencia, Spain

Understanding how people behave when facing hazardous situations, how intrinsic and extrinsic factors influence the risk taking (RT) decision making process and to what extent it is possible to modify their reactions externally, are questions that have long interested academics and society in general. In the spheres, among others, of Occupational Safety and Health (OSH), the military, finance and sociology, this topic has multidisciplinary implications because we all constantly face RT situations. Researchers have hitherto assessed RT profiles by conducting questionnaires prior to and after the presentation of stimuli; however, this can lead to the production of biased, non-realistic, RT profiles. This is due to the reflexive nature of choosing an answer in a questionnaire, which is remote from the reactive, emotional and impulsive decision making processes inherent to real, risky situations. One way to address this question is to exploit VR capabilities to generate immersive environments that recreate realistic seeming but simulated hazardous situations. We propose VR as the next-generation tool to study RT processes, taking advantage of the big four families of metrics which can provide objective assessment methods with high ecological validity: the real-world risks approach (high presence VR environments triggering real-world reactions), embodied interactions (more natural interactions eliciting more natural behaviors), stealth assessment (unnoticed real-time assessments offering efficient behavioral metrics) and physiological real-time measurement (physiological signals avoiding subjective bias). Additionally, VR can provide an invaluable tool, after the assessment phase, to train in skills related to RT due to its transferability to real-world situations.

Keywords: virtual reality, risk taking, occupational risks, risk attitude, risk perception, stealth assessment, psychophysiological assessment, embodiment

OPEN ACCESS

Edited by:
Samuel Greiff,
University of Luxembourg,
Luxembourg

Reviewed by:
Pedro Gamito,
Universidade Lusófona, Portugal
Valerie Shute,
Florida State University, United States

****Correspondence:***
Carla de-Juan-Ripoll
cardejua@i3b.upv.es

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: *01 September 2018*
Accepted: *27 November 2018*
Published: *12 December 2018*

Citation:
de-Juan-Ripoll C,
Soler-Domínguez JL, Guixeres J,
Contero M, Álvarez Gutiérrez N and
Alcañiz M (2018) Virtual Reality as
a New Approach for Risk Taking
Assessment. Front. Psychol. 9:2532.
doi: 10.3389/fpsyg.2018.02532

INTRODUCTION

Each year, deficient Occupational Safety and Health (OSH) practices cause a global cost of approximately 2680 billion euros (Elsler et al., 2017). Although OSH training has shown positive impacts in the workplace, its effectiveness is below expectations (Robson et al., 2012). It has been demonstrated that the natural differences between individuals can appreciably influence this low effectiveness at several levels, cognitive, motivational and functional, among others (Motowildo et al., 1997). Risk propensity, defined as the "willingness to take risks" (MacCrimmon and Wehrung, 1990) and risk perception, defined as the individual's assessment of how risky a situation is (Baird and Thomas, 1985), have been shown to have strong influence on risky decision making behaviors (Sitkin and Weingart, 1995). The measurement of risk taking (RT) attitudes is a recognized challenge for researchers and practitioners. Researchers have mostly employed self-report instruments to assess individual constructs based on theoretical psychological models (Brockhaus Sr, 1980; Ford et al., 1990; Gullone et al., 2000; Portell and Solé, 2001; Steinberg, 2004; Gardner and Steinberg, 2005; Sneddon et al., 2013; Rodríguez-Garzón et al., 2015). We have not found any one model that defines RT, thus its measurement requires further investigation. Lejuez et al. (2002) developed and validated a laboratory-based behavioral measure of RT (Balloon Analog Risk Task – BART). While this is a validated tool that has been used in several studies, we believe that it is desirable to develop a more ecological system to measure RT. VR provides the capability of creating interactive environments in which users can perform while their behavioral responses are recorded (Parsons, 2015). Accordingly, we propose that virtual environment based assessments are tools that can enhance the ecological validity of the evaluation of the responses evoked (Parsey and Schmitter-Edgecombe, 2013).

In this article we focus on the measurement of RT using physiological and behavioral metrics, with VR being employed as a tool to create immersive situations. We propose to use VR to assess RT attitudes under the paradigm of stealth assessment. VR can provide engaging virtual worlds which will allow real time measurement of RT behaviors.

This paper is comprised of four sections. In the first we review the theoretical framework of RT in the previous literature. In the second we summarize the extant instruments for the measurement of RT behaviors and discuss the current issues that make us believe that there is a need to establish a new approach. In the third we propose VR as a step forward in the assessment of RT. The fourth section briefly discusses the substantial implications raised by the article and our proposals for future research in this field.

RESEARCH INTO RISK TAKING

RT research can be said to have started with the nuclear debate of the sixties. It was focused on risk acceptance and dealt with factors such as benefits and voluntariness. Since then, several more factors have been proposed for the explanation of RT: trust, trustworthiness and trust propensity (Colquitt et al., 2007); supportive supervision, job autonomy and communication quality (Parker et al., 2001); problem framing and outcome history (Sitkin and Weingart, 1995); expected utility (Kahneman and Tversky, 1986); genre (Byrnes et al., 1999) and boredom (Schroeter et al., 2014).

While these factors have been demonstrated to influence RT, individual differences constitute a key element in decision making processes (see **Figure 1**). According to Rundmo, 1996, a biased perception of risk – understood as the subjective evaluation of a risk - can lead to misjudgements of potentially hazardous risk sources. Therefore, if the subjective evaluation of a risk differs from the objective risk, this should be corrected (Risk Research Committee, 1980). Personality traits influence attitude toward risk, prompting risk seeking or risk aversion behaviors. This set of personal, innate, basic characteristics associated with risk were named Intrinsic Risk Attitude (IRA) by Schoemaker (1993) and have been shown to be consistent in various situations and contexts (Dohmen et al., 2011). Additionally, cognitive and affective states are also considered to be key influencers in the decision making process. We highlight mood and cognitive load as two main representative factors in this category. Mood has a strong influence on RT. People in a positive mood tend to focus on the benefits of a risky situation, much more so than those in neutral mood, making them more susceptible to undertake risky behaviors (Forgas, 1982, 1995; Forgas and Bower, 1987; Yuen and Lee, 2003). On the other hand, people in a negative mood overestimate risks and try to avoid potential loss and, therefore, think and act more carefully (Jorgensen, 1996). Cognitive load, the amount of mental activity involved in working memory, might also play a role in risk perception, since some kind of decisions, based on utilitarian judgments, require additional cognitive resources (Greene et al., 2008).

RISK TAKING MEASURES: CURRENT ISSUES

RT measurement is a non-deterministic and non-standardized process based on different perspectives. Traditionally, most theories of human behavior are based on a model of the human mind that assumes that humans can think and verbalize accurately about their attitudes, emotions and behaviors (Simon, 1976; Brief, 1998). To date, most of the theoretical constructs used in RT assessment are based on explicit measures such as self-reports. However, recent advances in neuroscience have demonstrated that most of the brain processes that regulate our emotions, attitudes and behaviors are not conscious. That is, they are implicit processes that, in contrast to explicit processes, humans cannot verbalize (Barsade et al., 2009; George, 2009; Becker et al., 2011).

Several explicit measures of RT, oriented to evaluate attitude to risk, deferred risk perception or expected risk behavior, have been proposed in the last fifty years. Some authors have employed self-report measures based on questionnaires on compliance with safety practices in the workplace (Parker et al., 2001; Mohamed et al., 2009; Seo et al., 2015), attitude toward risk and organizational commitment (Kivimäki and Kalimo, 1993) and in studies into decision making (Sitkin and Weingart, 1995). On the other hand, some works have drawn on theoretical multidimensional models based on psychological constructs, such as personality (Lejuez et al., 2002; Skeel et al., 2007), impulsivity (Lejuez et al., 2002), sensation seeking (Horvath and Zuckerman, 1993; Lejuez et al., 2002) and situational awareness (Lejuez et al., 2002).

However, as in many other disciplines, pre- and post-experiment questionnaires have an important intrinsic bias

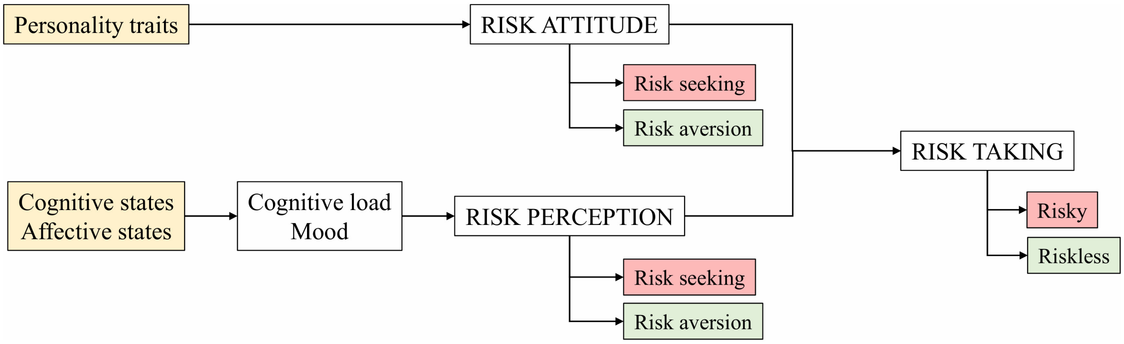


FIGURE 1 | Individual differences that influence risk taking.

TABLE 1 | VR features and benefits of risk taking measurement.

Domain	VR features	Benefits of measurement
Real-world risks	Evokes the sensation of physical risk	Neural mechanisms similar to real life
Embodied interactions	Actions raised in the first person	More emotional decisions
Stealth assessment	Indirect evaluation in real time	Reduction of test anxiety More validity and reliability
Physiological real-time measurement	Physiological measurement during performance	Involuntary, uncontaminated by participant answering bias

since individuals' cognitive and psychological states will be different when they answer the questionnaires to when they actually underwent the experiences that the researchers wish to analyse (Kivikangas et al., 2011). As stated in (Wang et al., 2015), this tendency is primarily due to "social desirability effects," which can lead to untrue accounts of behavior, attitudes and beliefs (Paulhus, 1991). In addition, there may be different interpretations of specific self-report items, resulting in unreliability and poorer validity (Lanyon and Goodstein, 1997). Lastly, some self-reporting questions need people to possess overt knowledge of their dispositions (Schmitt, 1994) and this does not always run true.

To our knowledge, the BART (Lejuez et al., 2002) constitutes, to date, the only tool for RT measurement using implicit measures. The authors developed and validated a laboratory-based behavioral measure of risky behaviors. In this task, a balloon was presented in the middle of the screen. Subjects were asked to pump it as much as possible, knowing that it could exploit at any time. Participants were told that they would obtain a financial reward the more they could inflate the balloon without breaking it. Although the reliability of this tool has been retested (White et al., 2008), extensive investigations have demonstrated that the correspondence between performance in neuropsychological tests and real-life behaviors is very weak (Manchester et al., 2004; Sbordone, 2008; Bottari et al., 2009).

In the BART validation study, researchers employed measures of impulsivity, sensation seeking and behavioral constraint. We consider this a good basis to build on, since each of these constructs has been investigated independently and associated with RT. Firstly, impulsivity has been associated with RT in terms of drug use, drink driving and seatbelt use (de Wit, 2009; Stanford et al., 1996). Some authors have also demonstrated its connection with emotional self-control, inhibition and, especially, the management of frustrating situations (Cooper et al., 2000; Boyer, 2006). In addition, researchers have studied the relationship between the sensation seeking trait and RT in several domains, such as recreation, health, career, finance, safety and social life (Nicholson et al., 2005). Donohew et al. (1999) concluded that sensation seeking is an important factor in sexual RT. According to Tellegen's (1985), model behavioral constraint is one of the dimensions that composes personality. The behavioral constraint factor encompasses control, harm avoidance and traditionalism facets. In the same way, there is empirical evidence of the influence of personality traits on RT attitudes, in particular punishment avoidance (Paulus et al., 2003). We can find an interesting study from Wills et al. (2006) supporting this idea in the substance abuse field.

LIMITATIONS OF CURRENT RISK TAKING MEASURES

As mentioned previously, to date the majority of RT assessment tools has been based on explicit measures and the use of questionnaires.

BART, with its multi-dimensional set of psycho-cognitive influences, represents the only alternative to explicit measures of RT behavior, but its design has some intrinsic limitations that current technologies could help to overcome.

In this regard, we believe that the existing measurement instruments do not reflect real situations, in which the subjects can perform as in real life, which leads to skewed results. In the laboratory the controlled stimuli given to subjects often do not include variables that are present in real life situations. Thus, the ecological validity of these methodologies, such as BART, is quite limited. Furthermore, these measurement tools do not involve any strong physical interaction, but require only simple actions, such as clicking a mouse, ignoring the influence of the reactions of the rest of the body. In addition, when an individual is submitted to the currently available tests, (s)he is aware that (s)he is being assessed and can alter the outcomes; so we propose stealth assessment as a means of obtaining reliable results about real behaviors unnoticed by the subject. Lastly, we suggest that physiological processes must be considered as important measures of RT, as these measurements are uncontaminated by the participant's answering style, social desirability, interpretations of questionnaire item wording, the limits of his or her memory or by observer bias (Kivikangas et al., 2011). Thus, we propose an alternative measurement method which aims to advance in four specific aspects:

(1) Real-world risks: As stated in Bornovalova et al. (2009), p.261. "[BART] ...... did not collect information on "real-world" risk-taking. It would be of both theoretical interest and clinical relevance to examine whether the current results "hold" when considering actual risk-taking behavior". We want to expose individuals to (almost) real risks in order to obtain (almost) real reactions. Amit et al. (2014) found that humans demonstrate two kinds of thought processes in any given situation, verbal and visual. A person who tends to verbal thinking builds meanings using words. This generates an abstract interpretation of a concept. It is usual, in this circumstance, to exhibit controlled cognitive processes, experience high psychological distance and to make utilitarian judgements. In contrast, visual thinking is associated with the use of images to represent concepts, generating a sense of proximity and the making of deontological judgements. People who tend toward visual thinking are willing to be guided by emotional automatic processes and are strongly influenced by secondary emotions. Using the real-world risks approach, we suggest that we can evoke the sensation of physical risk and initiate visual thinking that would arise in a real life, risky situation.

(2) Embodied cognition: How the actions of our bodies influence our perception, communication and learning processes is a field of study known as Embodied Cognition (EC). EC can be defined by stating that cognition is solidly based on corporal interactions with the physical environment (Wilson, 2002; Gallagher, 2005). Going into more detail, systems for sensing, acting and thinking are intrinsically interdependent and human cognition is made up of complex, specific representations combining all three systems (Soler et al., 2017). During recent years, instructional methods based on bodily interactions have been developed to create meaningful connections between physical activity and different knowledge domains, mainly in the STEM (Science, Technology, Engineering and Maths) area, strongly linked to the new Mixed Reality media (Lindgren and Johnson-Glenberg, 2013). To a certain extent, embodied learning could represent an important foundation on which to build a whole set of interactive, immersive learning environments. This concept is supported by previous research (Kontra et al., 2012) that argues that taking a meaningful action enhances learning in comparison to passively perceiving that action. This idea has been strongly supported for decades by classical learning theorists such as Piaget and Cook (1952) and Vygotsky (1978). We propose to take advantage of the ideas underlying embodied learning theory and use high level cognitive experiences, involving sensing, acting and thinking, to measure and change attitudes in a deeper, more effective way.

(3) Stealth assessment: "When embedded assessments are seamlessly woven into the fabric of the learning environment so that they are virtually invisible or unnoticed by the learner, this is stealth assessment" (Shute and Spector, 2008, unpublished, p.2). More specifically, this method offers the possibility of assessing different behaviors related to concrete capabilities, providing indirect evaluations in real time (Mislevy et al., 2003) and reducing test anxiety, while maintaining validity and reliability (Shute et al., 2008). Stealth assessment fits into the framework of evidence-centered design (ECD), which considers three conceptual models that must be present in stimuli design: the competency model, which aims to define the skills that the researcher wishes to assess; the evidence model, that aims to define specific behaviors and their relationships with particular skills and capabilities; and the task model, which is designed to develop specific scenarios and tasks to prompt skills-related behaviors (Shute, 2011). Thus, stealth assessment allows the setting of tasks and creation of situations that can elicit particular behaviors connected with the skills and capabilities to be evaluated.

(4) Physiological real-time measurement: Several physiological measures have recently been proposed as implicit measures of human behavior (Kivikangas et al., 2011). Skin conductance level has been successfully used as a measure of implicit processes such as stress, affective arousal and cognitive processing (Sequeira et al., 2009). Heart variability (HV) has been used for the implicit measurement of complex phenomena, for example cognitive load (Durantin et al., 2014). Eye tracking (ET) is a very interesting measure of subconscious brain processes, showing correlations with information processing in risky decisions (Glöckner and Herbold, 2011) and problem solving (Knoblich et al., 2001). Recent studies, using Functional Near-Infrared Spectroscopy (fNIRS), into decision making under pressure (Tsujii and Watanabe, 2010) and decision making processes in approach-avoidance theories (Ernst et al., 2013), are highly relevant for RT measures.

VIRTUAL REALITY AND RISK TAKING ASSESSMENT

Virtual Reality is a 3D synthetic environment able to simulate real experiences in which subjects can interact as if they were in the real world (Alcañiz et al., 2003). VR provides greater immersion, fidelity and higher level of active user involvement than traditional methods of assessment and training (Hedberg and Alexander, 1994). In our view, VR constitutes a suitable tool for behavioral measurement, since it complies with the requirements (see **Table 1**) of the four specific aspects discussed in the previous section: (1) the real-world risks approach, (2) embodied learning, (3) stealth assessment and (4) physiological real-time measurement.

(1) According to Slater (2009), the result of immersion through technology is the psychological state of "being there," where the subject essentially forgets that (s)he is in a virtual reality setting. This produces a sense of presence and a "plausibility illusion" which evoke the perception that what is happening in the VR is actual and allows subjects to interact and behave as they might in real life. VR is being used increasingly for natural phenomena and social interactions simulation, since it has been demonstrated that neural mechanisms in humans when they are immersed in a virtual environment are similar to those in real life (Alcañiz et al., 2009). When we talk about training and learning, failure is a necessary ingredient. There is evidence that people who have faced real hazards have a more cautious attitude toward OSH (Cavalcanti and Soares, 2012). Hazards in real life can involve serious danger. This is why VR emerges as a potential medium for RT assessment and training, allowing users to operate, without risks, in a quasi-real environment (Amokrane et al., 2008). VR allows the exposure of a person to a risky situation and the activation of high fidelity cognitive processes and behaviors due to the plausibility of the immersion. (2) VR environments allow users to take part in an embodied learning experience, mainly through physical interactions (Kilteni et al., 2012). Going further with this concept (Dourish, 1999, unpublished), we consider a virtual interaction to be fully embodied when it is believable, in the sense of using our body coherently as we do in the real world. The dual-process theory of moral judgment, when it refers to moral dilemmas, makes a distinction between personal and impersonal dilemmas (Greene et al., 2001; Greene, 2009): personal dilemmas are conflicts in which the subject experiences the situation in the first person and actions are carried out physically – e.g., pushing. Conversely, impersonal dilemmas are seen from the outside, and the subjects do not take overt physical actions, but make only minor responses, such as pressing switches or levers. Based on this distinction, it has been demonstrated that when actions are based on the first person perspective and involve physical acts, the subjects tend to make more emotional decisions (Greene et al., 2001; Amit et al., 2014). (3) Stealth assessment can be also defined as a performance-based method, in which what is evaluated is latent (Rupp et al., 2010). Under this paradigm, embedding assessments in immersive virtual worlds is an innovative approach (Shute and Spector, 2008) that, in our view, is an improvement from the standpoint of ecological validity. (4) Regarding physiological real-time measurement, VR provides interactive and multimodal sensorial stimuli that provide unique advantages over other methodologies in neuroscientific investigation (Bohil et al., 2011). Thus, due to technological advances, researchers can now use accurate, affordable devices to obtain physiological measures which have been found to be more effective than self-reported measures as they (a) are not intrusive, (b) do no rely on participants' self-assessment of their emotional or cognitive experience, and (c) can detect changes in participants in real time. We have previous experience in combining VR technology with brain activity measures, and these results have shown that interactive virtual environments allow the measurement of emotional responses (Marín-Morales et al., 2018).

For these reasons, customizable, domain independent VR environments, in which individuals can, to a certain extent, act freely and react naturally to different risks or hazards, open to researchers an uncharted field of information about RT attitudes and behaviors. The set of these requirements may result in an application that includes a virtual environment, with a specific narrative that face the users with risky situations. This should be designed following stealth assessment methodology, and would allow physiological and behavioral measurement to provide information about individual decision making in the field of RT. We will show an example of how this tool might perform: the user could be in a virtual environment that consists in a path which (s)he must cover from start to finish, within the shortest possible time. Suddenly, (s)he meets a bifurcation, where (s)he has to choose whether a safe but log way – less risk, less potential benefit -, or a dangerous but short path – higher risk, higher potential benefit -. During this decision making process, we could take measures of galvanic skin response to assess emotional activation, and behavioral measures such as reaction time and the decision made by the user. As a result, we could obtain information about specific weight of emotional processes in RT, and its influence on behavior.

Our future research aims to study to what extent a VR tool is able to measure the cognitive and affective processes that influence RT. Furthermore, we would focus on how virtual interactions and narratives weight on the decision making process.

CONCLUSION

RT measurement is a major challenge for companies and researchers. Investigations into behavioral measurement are at a turning point as, due to the potential of technological advances, we can generate virtual worlds to evaluate and, going further, train people in certain skills and competences. We suggest that virtual reality is the most appropriate medium for assessing attitudes to risk and risk perception, conditioning factors in the RT process, due to their immersive capabilities. We propose to undertake future investigations into real-world risks, embodied interactions, stealth assessment and physiological real-time measurement as differentiating elements in RT assessment. If we can study and measure the real, unbiased reactions of people facing risky or hazardous situations, it will be possible to create customized training programs to fit their individual characteristics. This can be expected to contribute to the improvement of OSH training programs, reducing work-related incidents and, consequently, costs for companies.

AUTHOR CONTRIBUTIONS

MA, CdJ, and NÁ were responsible for the general idea of the paper. CdJ and JS participated in drafting the work, while JG and MC revised it in-depth and provided new ideas thanks to their previous experience. MA supervised the entire work, revised the manuscript and approved the final version to be submitted. All authors made substantial contributions to the conception and development of the work.

FUNDING

This work was supported by the Spanish Ministry of Economy, Industry and Competitiveness funded projects "Advanced Therapeutic Tools for Mental Health" (DPI2016-77396-R), and "Assessment and Training on Decision Making in Risk Environments" (RTC-2017-6523-6) (MINECO/AEI/FEDER,UE).

REFERENCES

Alcañiz, M., Lozano, J. A., and Rey, B. (2003). Technological background of VR. *Stud. Health Technol. Inform.* 99, 199–214.

Alcañiz, M., Rey, B., Tembl, J., and Parkhutik, V. (2009). A neuroscience approach to virtual reality experience using transcranial doppler monitoring. *Presence Tel. Virtual Environ.* 18, 97–111. doi: 10.1162/pres.18.2.97

Amit, E., Gottlieb, S., and Greene, J. D. (2014). "Visual versus verbal thinking and dual-process moral cognition," in *Dual-Process Theories of the Social Mind*, eds J. W. Sherman, B. Gawronski and Y. Trope (New York, FL: Guilford Press), 340–354.

Amokrane, K., Lourdeaux, D., and Burkhardt, J. M. (2008). HERA: learner tracking in a virtual environment. *IJVR* 7, 23–30.

Baird, I. S., and Thomas, H. (1985). Toward a contingency model of strategic risk taking. *Acad. Manag. Rev.* 10, 230–243. doi: 10.5465/amr.1985.4278108

Barsade, S. G., Ramarajan, L., and Westen, D. (2009). Implicit affect in organizations. *Res. Organ. Behav.* 29, 135–162. doi: 10.1016/j.riob.2009.06.008

Becker, W. J., Cropanzano, R., and Sanfey, A. G. (2011). Organizational neuroscience: taking organizational theory inside the neural black box. *J. Manag.* 37, 933–961. doi: 10.1177/0149206311398955

Bohil, C. J., Alicea, B., and Biocca, F. A. (2011). Virtual reality in neuroscience research and therapy. *Nat. Rev. Neurosci.* 12, 752–762. doi: 10.1038/nrn3122

Bornovalova, M. A., Cashman-Rolls, A., O'donnell, J. M., Ettinger, K., Richards, J. B., and Lejuez, C. W. (2009). Risk taking differences on a behavioral task as a function of potential reward/loss magnitude and individual differences in impulsivity and sensation seeking. *Pharmacol. Biochem. Behav.* 93, 258–262. doi: 10.1016/j.pbb.2008.10.023

Bottari, C., Dassa, C., Rainville, C., and Dutil, E. (2009). The factorial validity and internal consistency of the instrumental activities of daily living profile in individuals with a traumatic brain injury. *Neuropsychol. Rehabil.* 19, 177–207. doi: 10.1080/09602010802188435

Boyer, T. W. (2006). The development of risk-taking: a multi-perspective review. *Dev. Rev.* 26, 291–345. doi: 10.1016/j.dr.2006.05.002

Brief, A. P. (1998). *Attitudes in and Around Organizations*. Thousand Oaks, CA: Sage.

Brockhaus, Sr. (1980). Risk taking propensity of entrepreneurs. *Acad. Manag. J.* 23, 509–520.

Byrnes, J. P., Miller, D. C., and Schafer, W. D. (1999). Gender differences in risk taking: a meta-analysis. *Psychol. Bull.* 125:367. doi: 10.1037/0033-2909.125.3.367

Cavalcanti, J., and Soares, M. (2012). Ergonomic analysis of safety signs: a focus of informational and cultural ergonomics. *Work* 41(Suppl. 1), 3427–3432. doi: 10.3233/WOR-2012-0619-3427

Colquitt, J. A., Scott, B. A., and LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. *J. Appl. Psychol.* 92, 909–927. doi: 10.1037/0021-9010.92.4.909

Cooper, M. L., Agocha, V. B., and Sheldon, M. S. (2000). A motivational perspective on risky behaviors: the role of personality and affect regulatory processes. *J. Pers.* 68, 1059–1088. doi: 10.1111/1467-6494.00126

de Wit, H. (2009). Impulsivity as a determinant and consequence of drug use: a review of underlying processes. *Addict. Biol.* 14, 22–31. doi: 10.1111/j.1369-1600.2008.00129

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: measurement, determinants, and behavioral consequences. *J. Eur. Econ. Assoc.* 9, 522–550. doi: 10.1111/j.1542-4774.2011.01015

Donohew, R. L., Hoyle, R. H., Clayton, R. R., Skinner, W. F., Colon, S. E., and Rice, R. E. (1999). Sensation seeking and drug use by adolescents and their friends: models for marijuana and alcohol. *J. Stud. Alcohol.* 60, 622–631. doi: 10.15288/jsa.1999.60.622

Durantin, G., Gagnon, J. F., Tremblay, S., and Dehais, F. (2014). Using near infrared spectroscopy and heart rate variability to detect mental overload. *Behav. Brain Res.* 259, 16–23. doi: 10.1016/j.bbr.2013.10.042

Elsler, D., Takala, J., and Remes, J. (2017). *An International Comparison of the Cost of Work-related Accidents and Illnesses*. Available at: https://osha.europa.eu/sites/default/files/publications/documents/international_comparison-of_costs_work_related_accidents.pdf

Ernst, L. H., Plichta, M. M., Lutz, E., Zesewitz, A. K., Tupak, S. V., Dresler, T., et al. (2013). Prefrontal activation patterns of automatic and regulated approach-avoidance reactions—a functional near-infrared spec- troscopy (fNIRS) study. *Cortex* 49, 131–142. doi: 10.1016/j.cortex.2011.09.013

Ford, M., Wentzel, K., Wood, D., Stevens, E., and Siesfeld, G. A. (1990). Processes associated with integrative social competence: emotional and contextual influences on adolescent social responsibility. *J. Adolesc. Res.* 4, 405–425. doi: 10.1177/074355488944002

Forgas, J. P. (1982). Reactions to life dilemmas: risk taking, success and responsibility attribution. *Aust. J. Psychol.* 34, 25–35. doi: 10.1080/00049538208254714

Forgas, J. P. (1995). Mood and judgment: the affect infusion model (AIM). *Psychol. Bull.* 117, 39–66. doi: 10.1037/0033-2909.117.1.39

Forgas, J. P., and Bower, G. H. (1987). Mood effects on person-perception judgments. *J. Pers. Soc. Psychol.* 53, 53–60. doi: 10.1037//0022-3514.53.1.53

Gallagher, S. (2005). *How the Body Shapes the Mind*. Oxford: Oxford University Press. doi: 10.1093/0199271941.001.0001

Gardner, M., and Steinberg, L. (2005). Peer influence on risk taking, risk preference, and risky decision making in adolescence and adulthood: an experimental study. *Dev. Psychol.* 41, 625–635. doi: 10.1037/0012-1649.41.4.625

George, J. M. (2009). The illusion of will in organizational behavior research: nonconscious processes and job design. *J. Manag.* 35, 1318–1339. doi: 10.1177/0149206309346337

Glöckner, A., and Herbold, A. K. (2011). An eye-tracking study on information processing in risky decisions: evidence for compensatory strategies based on automatic processes. *J. Behav. Decis. Mak.* 24, 71–98. doi: 10.1002/bdm.684

Greene, J. D. (2009). Dual-process morality and the personal/impersonal distinction: a reply to mcguire, langdon, coltheart, and mackenzie. *J. Exp. Soc. Psychol.* 45, 581–584. doi: 10.1016/j.jesp.2009.01.003

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., and Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition* 107, 1144–1154. doi: 10.1016/j.cognition.2007.11.004

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science* 293, 2105–2108. doi: 10.1126/science.1062872

Gullone, E., Moore, S., Moss, S., and Boyd, C. (2000). The adolescent risk-taking questionnaire: development and psychometric evaluation. *J. Adolesc. Res.* 15, 231–250. doi: 10.1177/0743558400152003

Hedberg, J., and Alexander, S. (1994). Virtual reality in education: defining researchable issues. *Educ. Media Int.* 31, 214–220. doi: 10.1080/0952398940310402

Horvath, P., and Zuckerman, M. (1993). Sensation seeking, risk appraisal, and risky behavior. *Personal. Individ. Differ.* 14, 41–52. doi: 10.1016/0191-8869(93)90173-Z

Jorgensen, P. F. (1996). "Affect, persuasion, and communication processes," in *Handbook of Communication and Emotion*, eds P. A. Andersen and L. K. Guerrero (Amsterdam: Elsevier), 403–422. doi: 10.1016/B978-012057770-5/50017-5

Kahneman, D., and Tversky, A. (1986). "Choices, values, and frames", in *Judgement and Decision Making: An Interdisciplinary Reader*, eds H. R. Arkes and K. R. Hammond (New York, NY: Cambridge University Press), 194–210.

Kilteni, K., Groten, R., and Slater, M. (2012). The sense of embodiment in virtual reality. *Presence Teleoperators Virtual Environ.* 21, 373–387. doi: 10.1162/PRES_a_00124

Kivikangas, J. M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Järvelä, S., et al. (2011). A review of the use of psychophysiological methods in game research. *J. Gaming Virtual Worlds* 3, 181–199. doi: 10.1386/jgvw.3.3.181_1

Kivimäki, M., and Kalimo, R. (1993). Risk perception among nuclear power plant personnel: a survey. *Risk Anal.* 13, 421–424. doi: 10.1111/j.1539-6924.1993.tb00742

Knoblich, G., Ohlsson, S., and Raney, G. E. (2001). An eye movement study of insight problem solving. *Mem. Cogn.* 29, 1000–1009. doi: 10.1016/j.actpsy.2008.08.008

Kontra, C., Goldin-Meadow, S., and Beilock, S. L. (2012). Embodied learning across the life span. *Topics Cogn. Sci.* 4, 731–739. doi: 10.1111/j.1756-8765.2012.01221

Lanyon, I., and Goodstein, L. D. (1997). *Personality Assessment*, 3rd Edn. New York, NY: Wiley.

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., et al. (2002). Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). *J. Exp. Psychol. Appl.* 8, 75–84. doi: 10.1037/1076-898X.8.2.75

Lindgren, R., and Johnson-Glenberg, M. (2013). Emboldened by embodiment: six precepts for research on embodied learning and mixed reality. *Educ. Res.* 42, 445–452. doi: 10.3102/0013189X13511661

MacCrimmon, K. R., and Wehrung, D. A. (1990). Characteristics of risk taking executives. *Manag. Sci.* 36, 422–435. doi: 10.1287/mnsc.36.4.422

Manchester, D., Priestley, N., and Jackson, H. (2004). The assessment of executive functions: coming out of the office. *Brain Injury* 18, 1067–1081. doi: 10.1080/02699050410001672387

Marín-Morales, J., Higuera-Trujillo, J. L., Greco, A., Guixeres, J., Llinares, C., Scilingo, E. P., et al. (2018). Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors. *Sci. Rep.* 8:13657. doi: 10.1038/s41598-018-32063-4

Mislevy, R. J., Almond, R. G., and Lukas, J. F. (2003). *A Brief Introduction to Evidence-Centered Design. ETS Research Report Series*. Princeton: Educational Testing Service, doi: 10.1002/j.2333-8504.2003.tb01908

Mohamed, S., Ali, T. H., and Tam, W. Y. V. (2009). National culture and safe work behaviour of construction workers in Pakistan. *Safety Sci.* 47, 29–35. doi: 10.1016/j.ssci.2008.01.003

Motowildo, S. J., Borman, W. C., and Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Hum. Perform.* 10, 71–83. doi: 10.1207/s15327043hup1002_1

Nicholson, N., Soane, E., Fenton-O'Creevy, M., and Willman, P. (2005). Personality and domain-specific risk taking. *J. Risk Res.* 8, 157–176. doi: 10.1080/1366987032000123856

Parker, S. K., Axtell, C. M., and Turner, N. (2001). Designing a safer workplace: importance of job autonomy, communication quality, and supportive supervisors. *J. Occup. Health Psychol.* 6, 211–228. doi: 10.1037/1076-8998.6.3.211

Parsey, C. M., and Schmitter-Edgecombe, M. (2013). Applications of technology in neuropsychological assessment. *Clin. Neuropsychol.* 27, 1328–1361. doi: 10.1080/13854046.2013.834971

Parsons, T. D. (2015). Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Front. Hum. Neurosci.* 9:660. doi: 10.3389/fnhum.2015.00660

Paulhus, D. L. (1991). "Measurement and control of response bias," in *Measures of Personality and Social Psychological Attitudes*, eds J. P. Robinson, P. R. Shaver, and L. S. Wrightsman (Cambridge, MA: Academic Press), 1759. doi: 10.1016/B978-0-12-590241-0.50006-X

Paulus, M. P., Rogalsky, C., Simmons, A., Feinstein, J. S., and Stein, M. B. (2003). Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism. *Neuroimage* 19, 1439–1448. doi: 10.1016/S1053-8119(03)00251-9

Piaget, J., and Cook, M. T. (1952). *The Origins of Intelligence in Children*, Vol. 8, (New York, NY: International University Press), 18.

Portell, M., and Solé, M. D. (2001). *Riesgo Percibido: Un Procedimiento de Evaluación. Disponible en la Red.* Available at: http://www.insht.es/InshtWeb/Contenidos/Documentacion/FichasTecnicas/NTP/Ficheros/501a600/ntp_578.pdf

Risk Research Committee (1980). *Accidents in Norway. How Do We Perceive and Handle Risk?* Oslo: Risk Research Committee.

Robson, L. S., Stephenson, C. M., Schulte, P. A., Amick, B. C. III, Irvin, E. L., Eggerth, D. E., et al. (2012). A systematic review of the effectiveness of occupational health and safety training. *Scand. J. Work Environ. Health* 38, 193–208. doi: 10.5271/sjweh.3259

Rodríguez-Garzón, I., Delgado-Padial, A., Martinez-Fiestas, M., and Lucas-Ruiz, V. (2015). The delay of consequences and perceived risk: an analysis from the workers' view point. *Rev. Facultad Ingeniería Univ. Antioquia* 74, 165–176.

Rundmo, T. (1996). Associations between risk perception and safety. *Safety Sci.* 24, 197–209. doi: 10.1016/S0925-7535(97)00038-6

Rupp, A. A., Gushta, M., Mislevy, R. J., and Shaffer, D. W. (2010). Evidence-centered design of epistemic games: measurement principles for complex learning environments. *J. Technol. Learn. Assess.* 8, 1–47.

Sbordone, R. J. (2008). Ecological validity of neuropsychological testing: critical issues. *Neuropsychol. Handb.* 367:394.

Schmitt, N. (1994). Method bias: The importance of theory and measurement. *J. Organ. Behav.* 15, 393–398. doi: 10.1002/job.4030150504

Schoemaker, P. J. (1993). Determinants of risk-taking: behavioral and economic views. *J. Risk Uncertain.* 6, 49–73. doi: 10.1007/BF01065350

Schroeter, R., Oxtoby, J., and Johnson, D. (2014). "AR and gamification concepts to reduce driver boredom and risk taking behaviours," in *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Seattle, WA, doi: 10.1145/2667317.2667415

Seo, H. C., Lee, Y. S., Kim, J. J., and Jee, N. Y. (2015). Analyzing safety behaviors of temporary construction workers using structural equation modeling. *Safety Sci.* 77, 160–168. doi: 10.1016/j.ssci.2015.03.010

Sequeira, H., Hot, P., Silvert, L., and Delplanque, S. (2009). Electrical autonomic correlates of emotion. *Int. J. Psychophysiol.* 71, 50–56. doi: 10.1016/j.ijpsycho.2008.07.009

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games Instruction* 55, 503–524.

Shute, V. J., Hansen, E. G., and Almond, R. G. (2008). You can't fatten a hog by weighing it–or can you? evaluating an assessment for learning system called ACED. *Int. J. Artif. Intell. Educ.* 18, 289–316.

Simon, H. A. (1976). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*. New York, NY: Macmillan.

Sitkin, S. B., and Weingart, L. R. (1995). Determinants of risky decision-making behavior: a test of the mediating role of risk perceptions and propensity. *Acad. Manag. J.* 38, 1573–1592. doi: 10.2307/256844

Skeel, R. L., Neudecker, J., Pilarski, C., and Pytlak, K. (2007). The utility of personality variables and behaviorally-based measures in the prediction of risk-taking behavior. *Personal. Individ. Differ.* 43, 203–214. doi: 10.1016/j.paid.2006.11.025

Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philos. Trans. R. Soc. B Biol. Sci.* 364, 3549–3557. doi: 10.1098/rstb.2009.0138

Sneddon, A., Mearns, K., and Flin, R. (2013). Stress, fatigue, situation awareness and safety in offshore drilling crews. *Safety Sci.* 56, 80–88. doi: 10.1016/j.ssci.2012.05.027

Soler, J. L., Contero, M., and Alcañiz, M. (2017). "VR serious game design based on embodied cognition theory," in *Proceedings of the Joint International Conference on Serious Games*, (Berlin: Springer), 12–21. doi: 10.1007/978-3-319-70111-0_2

Stanford, M. S., Greve, K. W., Boudreaux, J. K., Mathias, C. W., and Brumbelow, J. L. (1996). Impulsiveness and risk-taking behavior: Comparison of high-school and college students using the Barratt Impulsiveness Scale. *Personal. Individ. Differ.* 21, 1073–1075. doi: 10.1016/S0191-8869(96)00151-1

Steinberg, L. (2004). Risk taking in adolescence: what changes, and why? *Ann. N. Y. Acad. Sci.* 1021, 51–58. doi: 10.1196/annals.1308.005

Tellegen, A. (1985). "Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report," in *Anxiety and the Anxiety Disorders*, eds A. H. Tuma and J. D. Maser (Hillsdale, NJ: Lawrence Erlbaum Associates, Inc), 681–706.

Tsujii, T., and Watanabe, S. (2010). Neural correlates of belief-bias reasoning under time pressure: a near-infrared spectroscopy study. *Neuroimage* 50, 1320–1326. doi: 10.1016/j.neuroimage.2010.01.026

Vygotsky, L. (1978). Interaction between learning and development. *Read. Dev. Children* 23, 34–41.

Wang, L., Shute, V., and Moore, G. (2015). Lessons learned and best practices of stealth assessments. *Int. J. Gaming Computer Mediat. Simul.* 74, 66–87. doi: 10.4018/IJGCMS.2015100104

White, T. L., Lejuez, C. W., and de Wit, H. (2008). Test-retest characteristics of the Balloon Analogue Risk Task (BART). *Exp. Clin. Psychopharmacol.* 16, 565. doi: 10.1037/a0014083

Wills, T. A., Walker, C., Mendoza, D., and Ainette, M. G. (2006). Behavioral and emotional self-control: relations to substance use in samples of middle and high school students. *Psychol. Addict. Behav.* 20, 265–278. doi: 10.1037/0893-164X.20.3.265

Wilson, M. (2002). Six views of embodied cognition. *Psychon. Bull. Rev.* 9, 625–636. doi: 10.3758/BF03196322

Yuen, K. S., and Lee, T. M. (2003). Could mood state affect risk-taking decisions? *J. Affect. Disord.* 75, 11–18. doi: 10.1016/S0165-0327(02)00022-8

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Education

ORIGINAL RESEARCH
published: 04 January 2019
doi: 10.3389/feduc.2018.00116

Automated Feedback Can Improve Hypothesis Quality

Karel A. Kroeze[1,2]*, *Stéphanie M. van den Berg*[2], *Ard W. Lazonder*[3], *Bernard P. Veldkamp*[2] *and Ton de Jong*[1]

[1] Department of Instructional Technology, University of Twente, Enschede, Netherlands, [2] Department of Research Methodology, Measurement and Data Analysis, University of Twente, Enschede, Netherlands, [3] Behavioural Science Institute, Radboud University, Nijmegen, Netherlands

Stating a hypothesis is one of the central processes in inquiry learning, and often forms the starting point of the inquiry process. We designed, implemented, and evaluated an automated parsing and feedback system that informed students about the quality of hypotheses they had created in an online tool, the hypothesis scratchpad. In two pilot studies in different domains ("supply and demand" from economics and "electrical circuits" from physics) we determined the parser's accuracy by comparing its judgments with those of human experts. A satisfactory to high accuracy was reached. In the main study (in the "electrical circuits" domain), students were assigned to one of two conditions: no feedback (control) and automated feedback. We found that the subset of students in the experimental condition who asked for automated feedback on their hypotheses were much more likely to create a syntactically correct hypothesis than students in either condition who did not ask for feedback.

Keywords: automated feedback, hypotheses, inquiry learning, context-free grammars, online learning environment

OPEN ACCESS

Edited by:
Samuel Greiff,
University of Luxembourg,
Luxembourg

Reviewed by:
Trude Nilsen,
University of Oslo, Norway
Jessica Andrews-Todd,
Educational Testing Service,
United States

****Correspondence:***
Karel A. Kroeze
k.a.kroeze@utwente.nl

Specialty section:
This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Education

Received: *14 September 2018*
Accepted: *11 December 2018*
Published: *04 January 2019*

Citation:
Kroeze KA, van den Berg SM, Lazonder AW, Veldkamp BP and de Jong T (2019) Automated Feedback Can Improve Hypothesis Quality. Front. Educ. 3:116. doi: 10.3389/feduc.2018.00116

INTRODUCTION

Active forms of learning are seen as key to acquiring deep conceptual knowledge, especially in science domains (Hake, 1998; Freeman et al., 2014). One of the active forms of learning is inquiry learning. Inquiry learning has been defined in many different ways with as its kernel that the method starts from questions for which students need to find answers [see e.g., (Prince and Felder, 2007)]. In the current work, we focus on one of the ways inquiry is used in instruction, namely "learning science by doing science": students are expected to form and test hypotheses by performing experiments and analyzing data. In following an inquiry cycle, students learn both science content and the scientific method. In this study, we focus on the practice of the scientific method, and in particular on the creation of hypotheses.

Most models of inquiry-based learning encompass an orientation and conceptualization phase that enables students to familiarize themselves with the topic of investigation. Common activities during orientation are studying background information and conducting a few explorative experiments with the equipment at hand. The intended outcome of these initial explorations is the formation of theories and ideas, formalized in hypotheses (Pedaste et al., 2015). Hypotheses are integral to the inquiry cycle: they direct students' attention to specific aspects of the research problem and, hence, facilitate experimental design and data interpretation (Klahr and Dunbar, 1988; Zimmerman, 2007). In a classic study, Tschirgi (1980) found that both children and adults design more conclusive experiments when trying to test a hypothesis that contradicts prior evidence. Hypothesis testing also increases the amount of domain knowledge students gain from an inquiry (Burns and Vollmeyer, 2002; Brod et al., 2018), which is probably due to the fact that hypotheses, regardless of their specificity and truth value, provide direction to students' inquiry process (Lazonder et al., 2009).

The importance of hypothesizing nevertheless stands in marked contrast with its occurrence in high school science classes. Research has consistently shown that inquiry is a complex process in which students make mistakes (Mulder et al., 2010). Specifically, students of all ages have problems in formulating hypotheses, particularly when they are unfamiliar with the topic of inquiry (Gijlers and de Jong, 2005; Mulder et al., 2010), and when experimental data is anomalous (Lazonder, 2014). As a consequence, few students generate hypotheses on their own account, and when they do, they often stick to a single hypothesis that is known to be true (i.e., confirmation bias) or formulate imprecise statements that cannot be tested in research. These natural tendencies demonstrate that unguided inquiry learning is likely to be ineffective (Mayer, 2004; Kirschner et al., 2006; de Jong and Lazonder, 2014). However, *guided* inquiry learning has been shown to compare favorably to both direct instruction (D'Angelo et al., 2014) and unguided inquiry learning (Furtak et al., 2012), and helps foster a deeper conceptual understanding (Alfieri et al., 2011).

Inspired by these positive findings we set out to design and evaluate a software scaffold that presented students with automatically generated feedback on the quality of their hypotheses.

THEORETICAL FRAMEWORK

Adaptive and Automated Scaffolding

Inquiry learning often takes place in virtual or remote laboratories and, to be successful, should be supplemented with guidance (de Jong and Lazonder, 2014). Furthermore, de Jong and Lazonder (2014) postulated that different types of students require different types of guidance. Recent work on differentiated guidance lends credence to this argument, finding a moderating effect of students' age (Lazonder and Harmsen, 2016) and prior knowledge (van Riesen et al., 2018) on learning activities and knowledge gains. Moreover, Furtak et al. (2012) showed teacher-led inquiry activities to be more effective than student-led inquiry, implying that teachers are effective suppliers of guidance. However, given that teachers' time is an increasingly valuable resource, several adaptive software agents have recently been developed to support teachers on specific tasks and that adapt the guidance to students' characteristics. While Belland et al. (2016) found no added effect of limited adaptive scaffolding over static scaffolding, intelligent tutoring systems (Nye et al., 2014), adaptive environments (Durlach and Ray, 2011; Vandewaetere et al., 2011), and automated feedback (Gerard et al., 2015, 2016) have all shown promising results. The common-sense conclusion appears to be that the more guidance is adapted to the individual student, the better the guidance—and thus the student—performs. Indeed, Pedaste et al. (2015) recently identified the development of "*virtual teacher assistants that analyse and respond to individual learners to create meaningful learning activities*" as one of the main challenges in the field.

Although adaptive and automated elements are increasingly common in online learning environments (e.g., Aleven et al., 2010; Lukasenko et al., 2010; Vandewaetere et al., 2011; Gerard et al., 2015, 2016; Ryoo and Linn, 2016), they have typically been designed and implemented for a single learning activity in a specific domain. The reason for this is simple; even adaptive guidance for a single well-defined learning task generally requires years of research and development. Data must be gathered and coded, models have to be trained and fitted, appropriate feedback has to be fine-tuned and a digital environment has to be developed. Each of these steps involves the input of experts from different fields; teachers, statisticians, educational researchers, and computer scientists. As a result, scaffolds in multi-domain environments such as Go-Lab (de Jong et al., 2014) and WISE (Linn et al., 2003) generally do not adapt to the individual student, nor can they automatically assess products or provide context-sensitive feedback. The hypothesis scaffold we describe and test in this paper aims to fill this gap.

We have been unable to find any existing literature on the automated scoring of and feedback on free-text hypotheses. In contrast, a variety of increasingly sophisticated natural language processing (NLP) techniques have been employed for automated essay scoring. However, the techniques applied to scoring essays typically require a large amount of training data, and even when training data is available they are unlikely to provide the level of detail on the underlying structure of hypotheses required to give meaningful feedback. Training data is not readily available for hypotheses, and would be expensive to gather (Shermis and Burstein, 2013).

Anjewierden et al. (2015) noted that the "language" of hypotheses is a subset of natural language with a specific structure. They suggested using a domain-specific list of variables and categorical values (the *lexicon*), in conjunction with a *grammar* of hypotheses. Together, the lexicon and grammar could be used to create a hypothesis parses that is robust, and can be adapted to different domains with relative ease. The work reported here attempts to implement such a context-free grammar.

Feedback

The *informative tutoring feedback* model [ITF, (Narciss, 2006, 2008)] distinguishes between *internal* feedback and *external* feedback, and a wide variety of feedback *types*. Internal feedback is provided by individual cognitive monitoring processes (Ifenthaler, 2011), external feedback can be provided by for example; teachers, peers, or automated scaffolds. Both types of feedback may conflict with or reinforce an *internal reference value*. Careful feedback design can help students regulate their learning process, particularly when internal and external feedback conflict (Narciss, 2008).

The function of feedback may be *cognitive*, *meta-cognitive,* or *motivational*, and a distinction can be made between *simple* (e.g., knowledge of performance, correct result) and *elaborated* (e.g., knowledge about task constraints, mistakes, and concepts) forms of feedback. These components broadly overlap with *outcome*, *corrective* and *explanatory* feedback types (e.g., Johnson and Priest, 2014). In a second-order meta-analysis on the effects of feedback, Hattie and Timperley (2007) prescribed that good feedback should set clear goals (*feed up*), inform the student of their progress (*feed back*), and provide steps to improve (*feed forward*). Finally, *immediate* feedback has been shown to give larger benefits than *delayed* feedback (Van der Kleij et al., 2015).

Research Goal and Context

This project is performed in the Go-Lab ecosystem (de Jong et al., 2014). Go-Lab is an online environment where teachers and authors can share online and remote laboratories (Labs) and scaffolding applications (Apps). Apps and Labs can, together with multimedia material, be combined to create Inquiry Learning Spaces (ILS), which can also be shared on the Go-Lab environment. **Figure 1** shows a screenshot of a typical ILS. This ILS is organized in six phases that follow an inquiry cycle (in this case; Orientation, Conceptualization, Investigation, Interpretation, Conclusion, and Discussion), and can be navigated freely.

The hypothesis scratchpad app [**Figure 2**; (Bollen and Sikken, 2018)] is used to support students with hypothesis generation. This study aimed to create an adaptive version of the hypothesis scratchpad that can scaffold the individual student in hypothesizing in any domain, with a minimum of set-up time for teachers. This new version will need to (1) identify mistakes in students' hypotheses, and (2) provide students with appropriate feedback to correct these mistakes. If the app achieves both of these goals, it will be a considerable step toward "*empowering science teachers using technology-enhanced scaffolding to improve inquiry learning*" (Pedaste et al., 2015).

DESIGN

For this project the hypothesis scratchpad currently available in Go-Lab has been extended. An automated feedback system was developed that can identify flaws in students' hypotheses and provide tailored feedback that enables students to correct their mistakes. The aim is to improve the quality of students' hypotheses.

The following sections will (1) describe the main components of hypotheses and the criteria used to assess them, (2) introduce the process of parsing hypotheses and applying criteria, (3) present the feedback given to students, and (4) formalize the outcome measures and statistical analyses used.

Criteria

Quinn and George (1975) were the first to formally define a set of criteria for evaluating hypotheses: (1) *it makes sense*; (2) *it is empirical,* a (partial) scientific relation; (3) *it is adequate*, a scientific relation between at least two variables; (4) *it is precise*—a qualified and/or quantified relation; and (5) *it states a test*, an explicit statement of a test. Subsequent research on hypothesis generation has broadly followed the same criteria, or a subset thereof. Van Joolingen and De Jong (1991, 1993) used a "syntax" and a "precision" measure, that correspond roughly with the "it makes sense" and "precise' criteria of Quinn and George. Mulder et al. (2010) used a "specificity" scale, using criteria comparable to those of Quinn and George.

Based on the criteria used by Quinn and George, and the measures used by Van Joolingen and de Jong, we developed a set of criteria that could be implemented in automated feedback. **Table 1** lists these criteria, providing a short explanation and examples from the electrical circuits domain for each criterion. In the automated feedback, the first two criteria are straightforward in that they rely on the presence of certain words. The remaining criteria are established using a context-free grammar parser, which is described in the next section.

Parser

To detect mistakes, the automated system needs to interpret hypotheses on the criteria listed in **Table 1**. Given the observation that hypotheses are a relatively structured subset of natural language (Anjewierden et al., 2015), we can define a *context-free grammar* [CFG, (Chomsky, 1956)] that covers all well-structured hypotheses.

CFGs can be used to define natural languages, and are ideally suited to define heavily structured languages [e.g., programming languages, (Chomsky, 1956)]. A CFG is comprised of a set of *production rules*. All the sentences that can be produced by the repeated application of these rules are the *formal language* of that grammar.

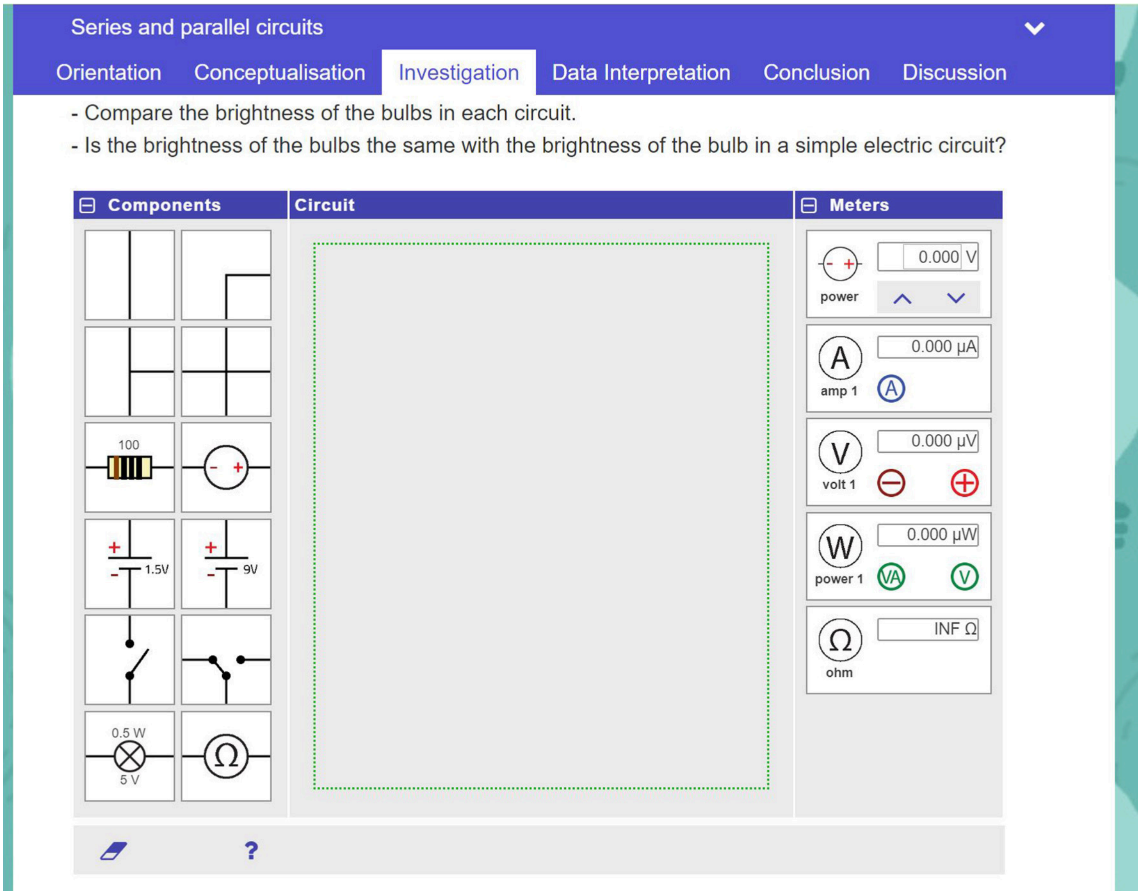
The grammar that defines hypotheses looks something like the following[1];

```
HYPOTHESIS -> if ACTION then ACTION
HYPOTHESIS -> ACTION if ACTION
ACTION -> VAR INTERACTOR VAR
ACTION -> VAR MODIFIER
ACTION -> MODIFIER VAR
ACTION -> ACTION and ACTION
VAR -> PROPERTY VAR
VAR -> bulbs
VAR -> voltage
VAR -> brightness
INTERACTOR -> is greater than
INTERACTOR -> is smaller than
INTERACTOR -> is equal to
MODIFIER -> increases
MODIFIER -> decreases
QUALIFIER -> series circuit
QUALIFIER -> parallel circuit
```

Each line is a production rule, the left-hand side of the rule can be replaced by the right-hand side. Uppercase words refer to further rules (they are *non-terminal*) and lowercase words refer to tokens (they are *terminal*). A token can be anything, but in our case, they are (sets of) words, e.g., "voltage" or "is greater than."

Consider the following hypothesis; *"if the number of bulbs in a series circuit increases, the brightness of the bulbs decreases."* If we were to apply our grammar, we can decompose this hypothesis

[1] For the complete grammar, see https://github.com/Karel-Kroeze/adaptive-hypothesis-grammars.



- Compare the brightness of the bulbs in each circuit.
- Is the brightness of the bulbs the same with the brightness of the bulb in a simple electric circuit?

FIGURE 1 | Screenshot of a typical inquiry learning space on the Go-Lab environment.

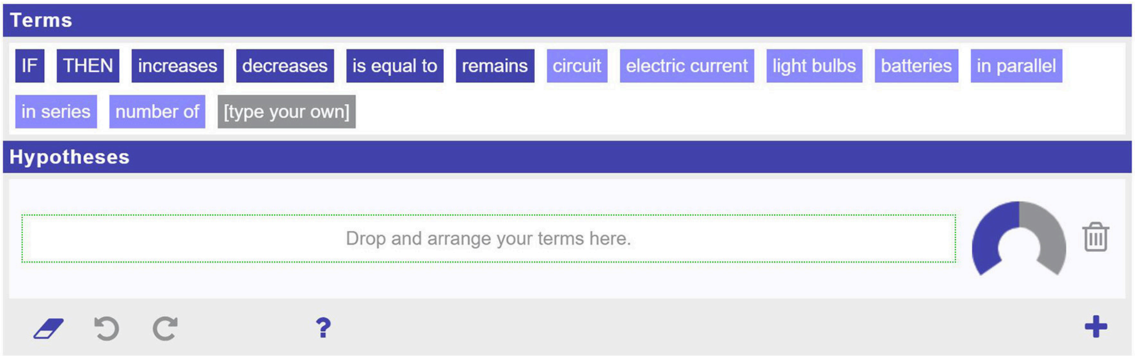


FIGURE 2 | Screenshot of the hypothesis scratchpad.

as per **Figure 3**. Although this decomposition provides the structure of the hypothesis, it still does not contain the *semantic* information necessary to evaluate the criteria.

If we add semantic information to each of the tokens, and rules on how to *unify* this information to each of the production rules, we can extract all relevant information from the hypothesis (Knuth, 1968; Shieber, 2003). **Figure 4** shows an example of the final parse result[2] which contains all the information needed to evaluate the criteria discussed.

[2] The parser was created using the Nearley.js package (Hardmath123., 2017), which implements the Earley context-free parsing algorithm (Earley, 1970). The source code of the parser is available on GitHub; https://github.com/Karel-Kroeze/adaptive-hypothesis-utils/.

TABLE 1 | Scoring criteria.

Criterion	Name	Description	Examples
1	Contains at least two variables	The hypothesis should contain at least two variables. Without two variables, the hypothesis can at best be an observation, and is likely to be nonsense.	✗"the current increases" ✓ "the current increases and the brightness increases"
2	Contains a modifier	The hypothesis should contain at least one modifier (e.g., "increases," "floats," but not "remains the same"). Without a modifier, the hypothesis can at best describe a static situation, and is likely to be nonsense.	✗ "the current remains the same" ✓ "the current increases"
3	Is a syntactically correct sentence	The hypothesis should be a correct sentence. Not only is the hypothesis likely to be nonsense if it is not a sentence, but moreover the automated system can only parse syntactically correct sentences.	✗ "the current increases decreases" ✓ "the current increases"
4	Manipulates exactly one independent variable	In order to test an effect of *x* on *y*, *x* should change, and no other variable should change.	✗ "if the current remains the same, the brightness increases" ✗ "if the number of bulbs increases and the current increases, the brightness remains the same" ✓ "if the number of bulbs increases, the brightness decreases"
5	Qualifies the variables	For some variables, it is their context that defines them. e.g., for buoyancy, density is defined by mass *and* volume, and in electrical circuits the *type* of circuit is crucial.	✗ "if the number of bulbs increases, the brightness remains the same" ✗ "if the mass of the object is larger than the volume of the fluid, the object sinks" ✓ "if the number of bulbs in a parallel circuit increases, the brightness remains the same"
6	Specifies interactions between variables	In some domains, it is the interaction between variables that is important. In our dataset this refers mainly to buoyancy, the relevant variable is the density of an object, as related to the density of the fluid.	✗ "if the density of the object increases, the object sinks" ✓ "if the density of the object is larger than the density of the fluid, the object sinks"

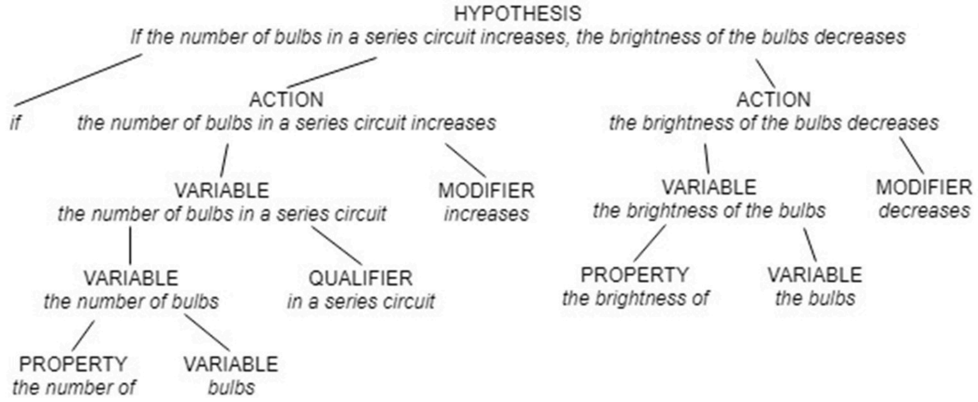

HYPOTHESIS
If the number of bulbs in a series circuit increases, the brightness of the bulbs decreases
if
ACTION
the number of bulbs in a series circuit increases
ACTION
the brightness of the bulbs decreases
VARIABLE
the number of bulbs in a series circuit
MODIFIER
increases
VARIABLE
the brightness of the bulbs
MODIFIER
decreases
VARIABLE
the number of bulbs
QUALIFIER
in a series circuit
PROPERTY
the brightness of
VARIABLE
the bulbs
PROPERTY
the number of
VARIABLE
bulbs

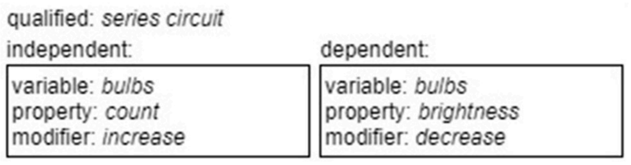
FIGURE 3 | Example of a hypothesis parse tree.

Feedback

The automated hypothesis scratchpad gives students the opportunity to request feedback. **Figure 5** shows an example of the automated hypothesis scratchpad, with the feedback button highlighted (the highlight is not part of the interface).

Table 2 gives an overview of the feedback used. The feedback follows the guidelines set by Hattie and Timperley (2007) in that it informs students of their progress, is specific about the mistakes made, and—where relevant—suggests modes of improvement. The first three criteria from **Table 2** are required conditions; if a hypothesis does not have variables, a modifier or cannot be parsed, the other criteria are not shown. Conversely, if these criteria are met, feedback is presented only on the other relevant criteria.

Feedback was presented to the student in textual form in a pop-up window and was shown immediately after a student



qualified: *series circuit*

independent:

variable: *bulbs*
property: *count*
modifier: *increase*

dependent:

variable: *bulbs*
property: *brightness*
modifier: *decrease*

FIGURE 4 | Parse result with semantic information.

requested it by clicking the feedback button. Feedback was never presented automatically. After receiving feedback, students could revise their hypothesis, and ask for feedback again. No explicit limits were placed on the amount of times students could ask for feedback.

Measures

Three outcome measures are of interest; (1) do students use the feedback tool, (2) does the parser correctly classify mistakes, and (3) do students' hypotheses improve after receiving feedback.

All student actions within a Go-Lab inquiry learning space are logged to a database. Specifically, the history of all hypotheses is tracked, including requests for feedback, and the feedback received. Feedback counts can thus be readily determined from the log files. A snapshot of a hypothesis is made whenever a student asks for feedback, and of the final state of the hypothesis. The collection of snapshots for a hypothesis creates a "story" for that hypothesis, tracking it over time.

The validity of classifications made by the parser is evaluated by calculating an inter-rater reliability between the results of the parser and human coders. The human coders were instructed to code as a teacher, ignoring small mistakes in spelling and syntax if the intention of a hypothesis was clear. To train the human coders, a sample of snapshots was coded, and any disagreements were discussed. After reaching agreement, each coder independently coded the remaining snapshots. Agreement is calculated using Cohens' κ, and interpreted using rules of thumb Landis and Koch (1977) .

Each snapshot is given a score based on the number of criteria passed, resulting in a score in a $0-k$ range, where k is the number of criteria used (three in the first pilot, six in the second pilot and final experiment). Improvement of hypotheses is evaluated by comparing the score for a snapshot to the score for the previous snapshot. The quality of a hypothesis is the quality of the final snapshot of that hypothesis.

If feedback is effective, we expect to see that students who have feedback available create higher quality hypotheses, and that hypothesis quality increases after students ask for feedback: each consecutive snapshot should have a higher quality than the last.

During the study, it became apparent that the aggregate score does not follow a parametric distribution, and therefore could not be used as an outcome measure. The variables and modifier criteria were satisfied by almost all students in our samples. The syntax criterion was often indicative for success on the manipulation, CVS and qualified criteria. Thus, even though the variables, modifier and CVS criteria might be important from a science education perspective, the syntactically correct criterion was used as an indicator for hypothesis quality.

Multilevel logistic models (i.e., generalized linear mixed models) were used to account for the inherent group structure in the data, controlling for student and class effects where appropriate. The models used were comprised of two levels, students and classes. All reported effects are on the student level. To perform the models, we used R (R Core Team, 2018) and the package "lme4" (Bates et al., 2015). The scripts used in analyses are deposited along with the raw and generated datasets at DANS (Kroeze, 2018).

FIELD STUDIES

Three field studies were conducted. An initial pilot study was conducted with an early version of the hypothesis parser to assess the feasibility of automated parsing of hypotheses using a context-free grammar. Following that, a second pilot study was conducted with the complete version of the parser to identify any remaining issues with the parser and ILS before moving on to the final experiment. The final experiment used a quasi-experimental design to assess the benefit of the tool in improving students' hypotheses. Each of these studies is described in more detail in the following sections.

First Pilot Study

Participants

Four classes of 13- to 14-year-old secondary education students ($n = 99$), spread over three HAVO classes (preparing for a university of applied science, $n = 76$) and one VMBO class (preparing for vocational education, $n = 23$) at a local high school participated in the pilot. Students had already studied the subject matter (supply and demand) as part of their regular curriculum and had previously participated in studies using Go-Lab ILSs and a version of the hypothesis scratchpad that did not provide feedback.

Materials and Procedure

The pilot revolved around a short ILS set in the *supply & demand* domain, where students were introduced to the interactions between price, supply, and demand. The ILS was created in collaboration with a participating economics teacher. Each class performed the study in a single 50-min session. At the beginning of a session, students were given an oral introduction detailing how to use the environment and refreshing them on what a hypothesis is. They were then asked to open the inquiry learning space, where they were first presented with information on the domain. They were then asked to create as many hypotheses about this domain as possible in the automated hypothesis scratchpad, and to use the feedback mechanism when they were stuck or wanted to check their hypothesis. An initial version of the parser was used that could detect the first three criteria: *it has* two *variables*, *it has a modifier*, and *it is a syntactically correct sentence*. Students were regularly encouraged and reminded to create as many hypotheses as

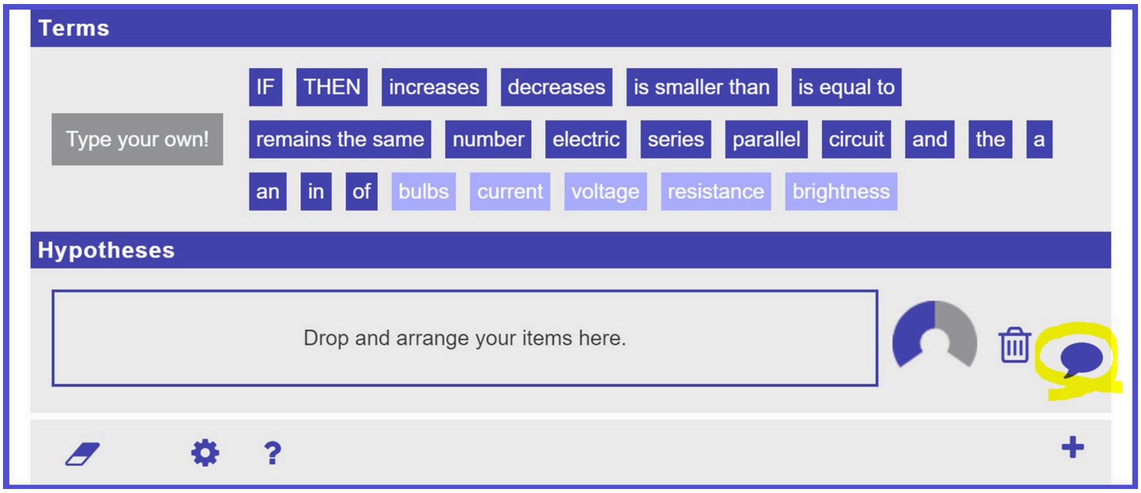


FIGURE 5 | Automated hypothesis scratchpad. The feedback button is highlighted.

TABLE 2 | Feedback for each criterion.

Criterion	Feedback	
	Wrong	**Correct**
Variables	Not enough variables, A hypothesis should always have at least two variables.	–
Modifier	You can only test a hypothesis if something changes. Without change, you cannot test the hypothesis.	–
Syntax	It appears you've entered an incomplete hypothesis. I can only give feedback on full hypotheses[a].	–
	I don't understand your hypothesis. Are you sure this is a correct hypothesis?; "[HYPOTHESIS]"[b]	–
CVS	If you don't change the value of [INDEPENDENT], you won't be able to test if this has an effect on [DEPENDENT][c]	You're changing the value of [INDEPENDENT] to see if that has an effect on [DEPENDENT][c]
	You're changing [INDEPENDENT] at the same time. You can't be sure which of these changes has an effect on [DEPENDENT][d]	You're changing only the value of [INDEPENDENT], so you can be certain that any change in [DEPENDENT] is caused by [INDEPENDENT][d]
Qualified	You did not describe the conditions in which your hypothesis applies.	You specified that your hypothesis only applies in a [QUALIFIER].

[HYPOTHESIS], [INDEPENDENT], [DEPENDENT], and [VARIABLE] will be dynamically replaced with the actual hypothesis and variables used by the student and recognized by the parser. The feedback has been translated from the Dutch original used in the experiments.

[a] *Used when a hypothesis starts valid but is incomplete (partial parse).*

[b] *Used when a hypothesis cannot be parsed (nonsense, or syntax error).*

[c] *Used when the independent variable is not manipulated.*

[d] *Used when multiple independent variables are manipulated.*

possible[3], but no attempt was made to force the creation of hypotheses or the use of the feedback tool. The session was concluded with a small user satisfaction questionnaire. During each session, the researcher and the classroom teacher monitored the class, answering process-related questions, and eliciting feedback if any out of the ordinary situations or interactions were encountered.

Results

A total of 979 hypotheses were collected from 96 students. Most students created three to five hypotheses and asked for feedback multiple times over the course of the experiment. One student asked for feedback 84 times and was removed as an outlier.

Inter-rater reliability between the parser and two human experts was almost perfect on all three criteria (Cohen's $\kappa = 0.81 - 0.96$), showing high parser accuracy. Hypotheses for which students requested and received feedback at least once were more likely to be correct on all criteria. This relation is visible in **Figure 6**, and statistically significant using a multilevel logistic model estimating the probability of a syntactically correct

[3] Unfortunately, during one of the HAVO sessions the teacher instructed students to create 'at least 4' hypotheses, which was immediately interpreted as 'create 4 hypotheses'.

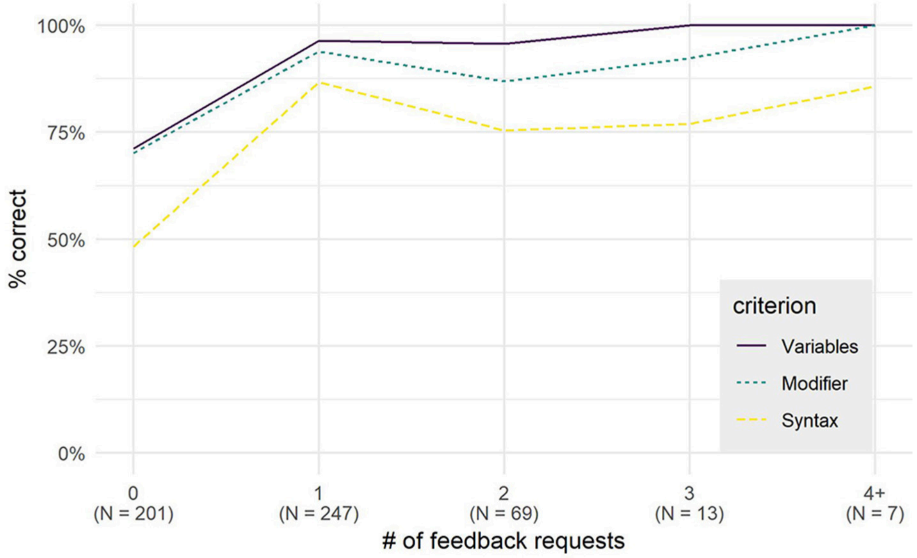


FIGURE 6 | Average performance on each criterion, by number of feedback requests.

hypothesis by the number of feedback requests, corrected for student and class effects, gender, and age ($\beta_{feedbackCount} = 1.00$, $SE_\beta = 0.17$, $CI_{OR} = 1.93 - 3.83$, $p < 0.001$), where $\beta_{feedbackCount}$ is the effect of each additional feedback request, and CI_{OR} the confidence interval of the Odds Ratio.

Discussion

The first pilot took place under test conditions; students were told to create as many hypotheses as possible, and the learning space was only there to provide a setting for hypotheses to be created. Such conditions are different from usual educational practice. Nevertheless, high parser accuracy and significantly increased quality of hypotheses showed that a parser is feasible, and that a hypothesis scratchpad enhanced with automated scoring and feedback is promising.

Therefore, a second pilot study was conducted using an expanded version of the context-free grammar that included all criteria listed in **Table 1**. In addition, the automated scratchpad was embedded in a full ILS, aligning much closer to how the tool is likely to be used in practice.

Second Pilot Study

Participants

Participants came from one HAVO class of 13 to 14-year-old secondary educations students ($n = 27$), at a local high school. The students had recently been introduced to electrical circuits as part of their regular curriculum but were familiar with neither Go-Lab environments nor the hypothesis scratchpad prior to the experiment.

Materials and Procedure

A short ILS in the *electrical circuits* domain that could be completed in a single 50-min session was created in collaboration with participating teachers. At the beginning of a session, students were given an oral introduction detailing how to use the tools in the ILS and refreshing them on what a hypothesis is. They were then asked to open the ILS, where they were presented with a short pre-test, followed by some information on the domain. To guide students' hypothesis construction, they were asked to enter two predictions about the change in brightness of lightbulbs in series and parallel circuits after adding another bulb. In the next steps, students were asked to turn these predictions into hypotheses in the automated hypothesis scratchpad, and design an experiment in the *Experiment Design* app [see e.g., (van Riesen et al., 2018)] to test their hypotheses. Finally, students were given time to create an experimental setup in the *Circuit Lab* virtual laboratory, test their hypotheses, and enter their conclusions.

All student actions took place in the ILS, which encompassed a full inquiry cycle, from orientation to conclusion. This created an environment more likely to occur in real educational settings. An expanded version of the automated hypothesis scratchpad was used, designed to be able to classify and give feedback to all the relevant criteria.

During the session, the researcher and the classroom teacher monitored the class, answering process-related questions and eliciting feedback if any out of the ordinary situations or interactions were encountered.

Results

Both the researcher and the classroom teacher noticed that students had problems working with the ILS and staying on-task. These problems were process related (e.g., students got distracted, skipped steps) and tool related (i.e., students did not know how to work with the tool). Attempts to provide instructions during the experiment were largely ineffective because students were at different stages of the ILS (making group instructions difficult), and there were too many students to provide individual instructions.

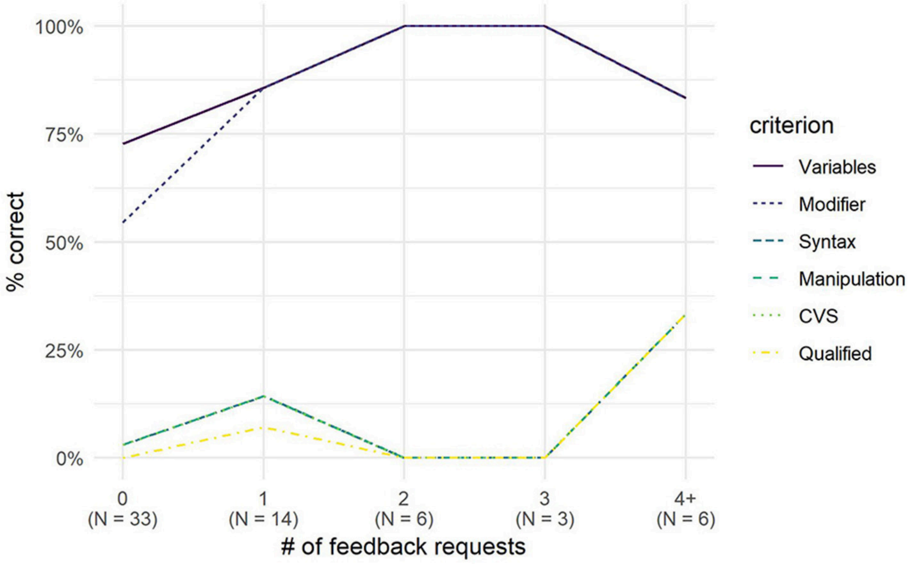


FIGURE 7 | Average performance on each criterion, by number of feedback requests. Note that the poor performance is at least partially due to low parser accuracy, and that the scores for the Syntax, manipulation, and qualified criteria overlap.

In addition, some of the written instructions were too long. For example, upon seeing the instructions, one student immediately uttered: "*too long, won't read.*" It seems likely that his sentiments were shared by other students, highlighting the need for verbal (or at least more interactive) instructions.

A total of 50 hypotheses were collected from 27 students. The plurality (13) of students created two hypotheses each, 7 students did not create any hypotheses. Most (16) students asked for feedback at least once, 11 students did not ask for feedback. One student asked for feedback 23 times and was removed as an outlier.

Parser accuracy was below expectations, achieving a Cohens' κ of 0.91, 0.90, and 0.40 on the *contains at least* two *variables*, *contains a modifier*, and *is a syntactically correct sentence* criterion, respectively. Accuracy for the *manipulates exactly* one *variable* and *is qualified* criteria is not reported, as the parser failed to recognize 30 out of 46 syntactically correct snapshots, leaving only 16 parsed snapshots.

Although there does appear to be a positive effect of feedback on hypothesis quality (see **Figure 7**), this effect was not statistically significant, as shown by a multilevel logistic model estimating the probability of a syntactically correct hypothesis by the number of feedback requests, correcting for student effects, gender and age ($\beta_{feedbackCount} = 0.46$, $SE_{\beta} = 0.24$, $CI_{OR} = 0.98 - 2.57$, $p = .058$).

Discussion

The number of collected hypotheses per student was lower than in the first pilot. In part, that was by design: the first pilot was specifically set up to encourage students to create as many hypotheses as possible, whereas, in this pilot students were guided to create two hypotheses. The participants in this pilot also had less experience working in an ILS, which caused several process-related issues during the session that likely influenced the number of hypotheses created. A more structured lesson plan where students start and end each step in the inquiry cycle at the same time will allow for verbal instructions to be given before starting each section.

Many students failed to distinguish between series and parallel circuits in their hypotheses, even when their predictions did show they understood the differences between the types of circuits. This does seem to indicate the need for supporting the creation of hypotheses while at the same time highlighting that the currently implemented support is insufficient.

Poor parser accuracy can be attributed to students' difficulties in working with the ILS, additional criteria introducing more complexity to the grammar, and a lack of training data for the *Electrical Circuits* domain in the target language (Dutch) to calibrate the parser. Using the data gathered in the pilot, we were able to make improvements to the grammar used by the parser. When applying this new grammar to the gathered hypotheses, inter-rater agreement on the syntax criterion was raised to moderate (Cohens' $\kappa = 0.53$).

Main Study

Participants

Six classes of 13- to 15-year-old secondary education students ($n = 132$), from two local high schools participated in the study. Six students used incorrect login credentials and were left out of the analyses. The remaining participants came from 4 HAVO classes ($n = 78$), and 2 VWO classes ($n = 48$). Students were randomly assigned to one of two conditions. Students in the experimental condition ($n = 68$) used the automated hypothesis scratchpad, while those in the control condition ($n = 58$) used a version of the hypothesis scratchpad that did not provide feedback. No significant differences were present in the distribution of age, gender, and current physics grade across conditions (**Table 3**).

Materials and procedure

A single 50-min session was used, covering the same material as that of the second pilot study. The ILS used in the second pilot study was used again, with some minor changes to ameliorate some of the process-related issues students encountered. In particular, written descriptions and instructions were shortened. Instead, at the outset of the session and each phase, students were given a short oral introduction.

Students received a link to a randomizer[4] that assigned each student to one of two conditions and redirected them to the corresponding ILS. Students were instructed not to move to the next phase until told to do so.

At pre-set intervals during the sessions, the researcher gave an oral introduction to the next phase of the inquiry cycle, and the corresponding tools in the ILS. Students where then encouraged to start with that phase. In each session, the researcher and the class teacher monitored the students, answering process-related questions, and eliciting feedback if any extra-ordinary situations or interactions were encountered.

Results

Most students were already familiar with the GoLab environment and its tools and encountered no significant difficulties. Based on observations during the sessions, oral introductions prior to each phase of the ILS appeared to keep most students on task, most of the time.

Students in the experimental condition created 201 hypotheses, for 56 of which feedback was requested. Of the 68 students in the experimental condition, exactly half never asked for feedback.

Parser accuracy was moderate to almost perfect, achieving a Cohens' κ of 0.84, 0.70, and 0.59 on the *contains at least* two *variables*, *contains a modifier*, and *is a syntactically correct sentence* criterion, respectively, and > 0.80 for the *manipulates exactly* one *variable* and *is qualified* criteria.

Figure 8 appears to show that on average the hypotheses generated in the experimental condition scored higher on all criteria. In addition, **Figure 9** suggests a positive relation between the number of feedback requests and the quality of hypotheses. In particular, hypotheses for which feedback was requested at least once appear to be of higher quality.

To test the effect of our tool on hypothesis quality, we fitted a multilevel logistic model, controlling for student and class effects, as well as gender, age, physics grade, and academic level. We found no significant effect from being assigned to the experimental condition ($\beta_{condition} = 0.25$, $SE_{\beta} = 0.34$, $CI_{OR} = 0.66 - 2.50$, $p = 0.472$). Given that half of all participants in the experimental group never requested feedback, this outcome was not unexpected.

However, when we split the experimental group in two, based on whether students requested feedback or not ($n = 34$ in both groups, **Figure 10**), and contrast those who requested feedback against those who did not or could not, controlling for student and class effects, as well as gender, age, physics grade and academic level, the effect of requesting feedback is significant ($\beta_{feedbackCount} = 1.47$, $SE_{\beta} = 0.42, CI_{OR} = 1.92 - 9.89$, $p < 0.001$).

It could be argued that students who did not request feedback when it was made available to them are less proficient students. However, a contrast analysis comparing students in the control condition (who could not ask for feedback) and those in the experimental condition who did not request feedback found no significant difference between the two groups on the syntactically correct criterion ($\beta_{condition} = -0.30$, $SE = 0.39$, $CI_{OR} = 0.34 - 1.60$, $p = 0.445$). We thus found no evidence to suggest that there was a difference between students who could have asked for feedback but did not do so, and students who did not have the option to ask for feedback.

GENERAL DISCUSSION

The creation of hypotheses is a critical step in the inquiry cycle (Zimmerman, 2007), yet students of all ages experience difficulties creating informative hypotheses (Mulder et al., 2010). Automated scaffolds can help students create informative hypotheses, but their implementation in the regular curriculum is often cost-prohibitive, especially since they can typically only be used in one specific domain and language. This study set out to create a hypothesis scratchpad that can automatically evaluate and score hypotheses and provide students with immediate feedback. We use a flexible Context-Free Grammar approach that can relatively easily be adapted and extended for other languages and domains. We described the development process of this tool over two pilot studies and evaluated its instructional effectiveness in a controlled experiment.

Across three studies, we showed that a hypothesis parser based on a context-free-grammar is feasible, attaining moderate to almost perfect levels of agreement with human coders. The required complexity of the parser is directly linked to the syntactical complexity of the domain. For example, the electrical circuits domain requires a more complex parser than the supply and demand domain. Further development of the context-free-grammar used in the parser will contribute to higher reliability and may extend it to other languages and domains.

The second pilot study illustrated that a lack of familiarity of students with the online environment and the tools used can have a negative effect on their performance. Students were distracted by technical and process related issues, and had difficulty remaining on-task. In the final experiment, we used a largely identical learning environment, but students were verbally introduced to each phase. These introductions allowed students to focus on the content of the learning environment, rather than on how to use the learning environment itself.

Nevertheless, when using the automated hypothesis scratchpad in a "typical" ILS, students often did not request

[4] A separate ILS was created for each condition. The randomizer forwarded the students browser to one of these conditions. Randomization was weighted to ensure a roughly equal distribution across conditions in each session.

TABLE 3 | Participant characteristics, by condition.

		Overall	Control	Experimental	Test statistic (df)	*P*-value
		126	58	68		
Gender (%)	Female	57 (45.2)	27 (46.6)	30 (44.1)	$\chi^2(1) = 0.01$	0.925
	Male	69 (54.8)	31 (53.4)	38 (55.9)		
Level (%)	HAVO	78 (61.9)	35 (60.3)	43 (63.2)	$\chi^2(1) = 0.02$	0.882
	VWO	48 (38.1)	23 (39.7)	25 (36.8)		
Mean age (SD)		13.96 (0.64)	13.98 (0.66)	13.95 (0.63)	$t(119.09) = 0.34$	0.736
Mean grade (SD)		6.50 (0.86)	6.52 (0.86)	6.49 (0.86)	$t(120.51) = 0.21$	0.836

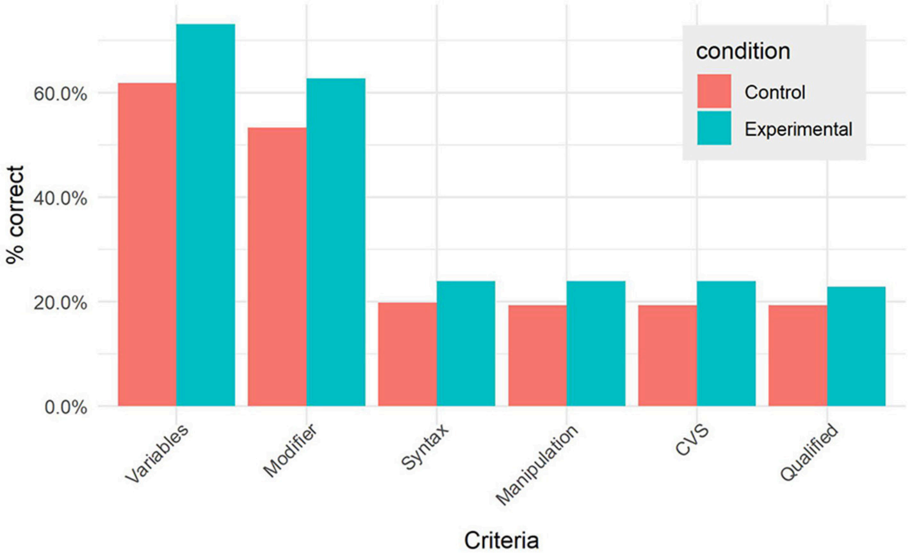


FIGURE 8 | Average performance by criterion, by condition.

feedback. Timmers et al. (2015) found a relation between gender and the willingness to ask for feedback, but such a relation was not present in our sample. In fact, none of the background variables collected (age, gender, physics grade and educational level) were significantly related to feedback requests or the quality of hypotheses.

If the goal was to obtain as many hypotheses as possible and assess the performance of the parser alone, we would have been better off following the approach taken in the first pilot. However, we deliberately chose to embed the automated hypothesis scratchpad in a typical ILS in the second pilot and main study, with the aim of replicating "real-world" conditions. In doing so, we can draw conclusions that are likely to be applicable to educational practice, rather than in laboratory conditions alone.

In the first pilot, the number of feedback requests was significantly related to the quality of hypotheses. This result was confirmed in a controlled experiment, where students who requested feedback were significantly more likely to create syntactically valid hypotheses than those who did not. The effects of feedback were immediate; hypotheses for which feedback was requested once where more likely to be correct.

To the best of our knowledge, no other tool exists that can reliably score hypotheses, can easily be adapted to different domains, and that allows students to create free-text hypotheses. The automated hypothesis scratchpad we present here can provide a clear and immediate benefit in science learning, provided students request feedback. By increasing the quality of students' hypotheses, we may assume that students are able to engage in more targeted inquiries, positively impacting their learning outcomes. How students can best be encouraged to request (and use) feedback is an open problem, and out of scope for this project. The automated hypothesis scratchpad could also be adapted to be a monitoring tool, highlighting students that may have difficulties creating hypotheses, allowing teachers to intervene directly.

The ability to reliably score hypotheses presents possibilities besides giving feedback. For example, hypothesis scores could serve as an indicator of inquiry skill. As such, they can be part of student models in adaptive inquiry learning environments. Crucially, obtaining an estimate from students' inquiry products

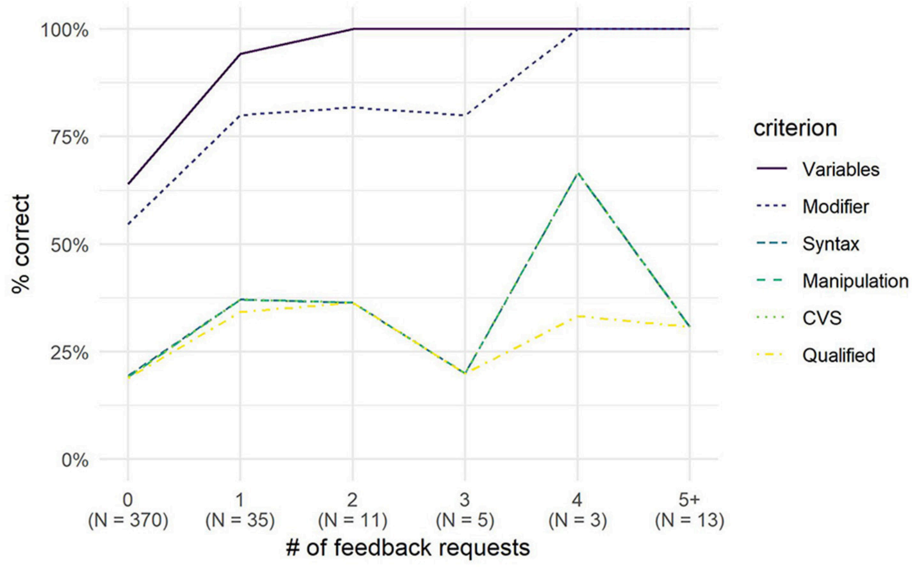


FIGURE 9 | Average performance by criterion for the experimental group, by number of feedback requests.

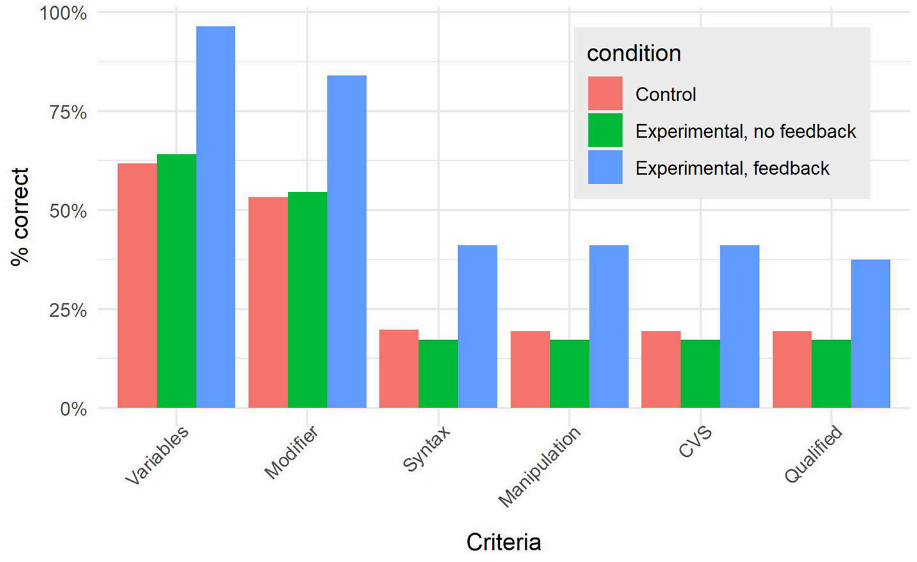


FIGURE 10 | Average performance on each criterion, by condition and feedback use.

is less obtrusive than doing so with a pre-test, and likely to be more reliable than estimates obtained from students' inquiry processes.

The aggregate hypothesis score computed for students did not have a known parametric distribution. This represents a serious limitation, as the score could not be used in statistical analyses. As a result, we chose to only test statistical significance based on the syntax criterion. Investigating alternative modeling techniques to arrive at a statistically valid conclusion based on multiple interdependent criteria will be part of our future work.

An automated hypothesis scratchpad providing students with immediate feedback on the quality of their hypotheses was implemented using context-free grammars. The automated scratchpad was shown to be effective; students who used its feedback function created better hypotheses than those who did not. The use of context-free grammars makes it relatively straightforward to separate the basic syntax of hypotheses, language specific constructs, and domain specific implementations. This separation allows for the quick adaptation of the tool to new languages and domains, allowing configuration by teachers, and inclusion in a broad range of inquiry environments.

ETHICAL STATEMENT

All participating schools have obtained written and informed consent from students' parents to perform research activities that fall within the regular curriculum. Parents were not asked to give consent for this study specifically. The experiments we performed were embedded in the students' curriculum, and the collected data was limited to learning processes and outcomes. Students were briefed that their activities in the online learning environment would be logged, and that this data would be used in anonymized form. Both the research protocol and consent procedures followed were approved by the ethical board of the faculty of Behavioural, Management and Social Sciences of the University of Twente (ref # 17029).

AUTHOR CONTRIBUTIONS

KK, TdJ, AL, and SvdB designed the experiment. KK and AL designed the intervention. KK, SvdB, and BV performed statistical analyses, TdJ and AL helped put experimental results into context. KK wrote the manuscript, aided by TdJ, SvdB, AL, and BV.

REFERENCES

Aleven, V., Roll, I., Mclaren, B. M., and Koedinger, K. R. (2010). Automated, unobtrusive, action-by-action assessment of self-regulation during learning with an intelligent tutoring system. *Educ. Psychol.* 45, 224–233. doi: 10.1080/00461520.2010.517740

Alfieri, L., Brooks, P., Aldrich, N. J., and Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? A meta-analysis. *J. Educ. Psychol.* 103, 1–18. doi: 10.1037/a0021017

Anjewierden, A., Kamp, E. T., and Bollen, L. (2015). *Analysis of Hypotheses in Go-Lab*. Enschede.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using {lme4}. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Belland, B. R., Walker, A. E., Kim, N. J., and Lefler, M. (2016). Synthesizing results from empirical research on computer-based scaffolding in STEM education: a meta-analysis. *Rev. Educ. Res.* 87, 309–344. doi: 10.3102/0034654316670999

Bollen, L., and Sikken, J. (2018). *Hypothesis Scratchpad*. Available online at: https://www.golabz.eu/app/hypothesis-scratchpad (Accessed November 28, 2018).

Brod, G., Hasselhorn, M., and Bunge, S. A. (2018). When generating a prediction boosts learning: the element of surprise. *Learn. Instr.* 55, 22–31. doi: 10.1016/j.learninstruc.2018.01.013

Burns, B. D., and Vollmeyer, R. (2002). Goal specificity effects on hypothesis testing in problem solving. *Q. J. Exp. Psychol. Sec. A* 55, 241–261. doi: 10.1080/02724980143000262

Chomsky, N. (1956). Three models for the description of language. *IRE Transac. Inform. Theory* 2, 113–124. doi: 10.1109/TIT.1956.1056813

D'Angelo, C., Rutstein, D., Harris, C., Bernard, R., Borokhovski, E., and Haertel, G. (2014). *Simulations for STEM Learning: Systematic Review and Meta-Analysis Executive Summary*. Menlo Park, CA: SRI International.

de Jong, T., and Lazonder, A. W. (2014). "The guided discovery learning principle in multimedia learning," in *The Cambridge Handbook of Multimedia Learning,* ed R. Mayer (Cambridge: Cambridge University Press), 371–390. doi: 10.1017/CBO9781139547369.019

de Jong, T., Sotiriou, S., and Gillet, D. (2014). Innovations in STEM education: the Go-Lab federation of online labs. *Smart Learn. Environ.* 1:3. doi: 10.1186/s40561-014-0003-6

Durlach, P. J., and Ray, J. M. (2011). *Designing Adaptive Instructional Environments: Insights From Empirical Evidence (Technical Report 1297).* Available online at: http://www.adlnet.gov/wp-content/uploads/2011/11/TR-1297.pdf

Earley, J. (1970). An efficient context-free parsing algorithm. *Commun. ACM* 13, 94–102. doi: 10.1145/362007.362035

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., et al. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proc. Natl. Acad. Sci.U.S.A.* 111, 8410–8415. doi: 10.1073/pnas.1319030111

Furtak, E. M., Seidel, T., Iverson, H., and Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: a meta-analysis. *Rev. Educ. Res.* 82, 300–329. doi: 10.3102/0034654312457206

Gerard, L. F., Matuk, C., McElhaney, K., and Linn, M. C. (2015). Automated, adaptive guidance for K-12 education. *Educ. Res. Rev.* 15, 41–58. doi: 10.1016/j.edurev.2015.04.001

Gerard, L. F., Ryoo, K., McElhaney, K. W., Liu, O. L., Rafferty, A. N., and Linn, M. C. (2016). Automated guidance for student inquiry. *J. Educ. Psychol.* 108, 60–81. doi: 10.1037/edu0000052

Gijlers, H., and de Jong, T. (2005). The relation between prior knowledge and students' collaborative discovery learning processes. *J. Res. Sci. Teach.* 42, 264–282. doi: 10.1002/tea.20056

Hake, R. R. (1998). Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. *Am. J. Phys.* 66, 64–74. doi: 10.1119/1.18809

Hardmath123. (2017). *Nearley.js. GitHub*. Available online at: https://nearley.js.org

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487

Ifenthaler, D. (2011). Bridging the gap between expert-novice differences: the model-based feedback approach. *J. Res. Technol. Educ.* 43, 103–117. doi: 10.1080/15391523.2010.10782564

Johnson, C. I., and Priest, H. A. (2014). "The feedback principle in multimedia learning," in *The Cambridge Handbook of Multimedia Learning*, ed R. Mayer (Cambridge: Cambridge University Press), 449–463. doi: 10.1017/CBO9781139547369.023

Kirschner, P. A., Sweller, J., and Clark, R. E. (2006). Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential and inquiry-based teaching. *Educ. Psychol.* 41, 75–86. doi: 10.1207/s15326985ep4102_1

Klahr, D., and Dunbar, K. (1988). Dual space search during scientific reasoning. *Cogn. Sci.* 12, 1–48. doi: 10.1207/s15516709cog1201_1

Knuth, D. E. (1968). Semantics of context-free languages. *Math. Syst. Theory* 2, 127–145. doi: 10.1007/BF01692511

Kroeze, K. A. (2018). *Automated Hypothesis Scratchpad*. DANS. doi: 10.17026/dans-znq-knky

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310

Lazonder, A. W. (2014). "Inquiry learning," in *Handbook of Research on Educational Communications and Technology,* eds J. M. Spector, M. D. Merrill, J. Elen, and M. J. Bishop (New York, NY: Springer New York), 1–34. doi: 10.1007/978-1-4614-3185-5_36

Lazonder, A. W., and Harmsen, R. (2016). Meta-analysis of inquiry-based learning: effects of guidance. *Rev. Educ. Res.* 86, 681–718. doi: 10.3102/0034654315627366

Lazonder, A. W., Wilhelm, P., and van Lieburg, E. (2009). Unraveling the influence of domain knowledge during simulation-based inquiry learning. *Instr. Sci.* 37, 437–451. doi: 10.1007/s11251-008-9055-8

Linn, M. C., Clark, D., and Slotta, J. D. (2003). WISE design for knowledge integration. *Sci. Educ.* 87, 517–538. doi: 10.1002/sce.10086

Lukasenko, R., Anohina-Naumeca, A., Vilkelis, M., and Grundspenkis, J. (2010). Feedback in the concept map based intelligent knowledge assessment system. *Sci. J. Riga Technic. Univ. Comp. Sci.* 41, 8–15. doi: 10.2478/v10143-010-0020-z

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *Am. Psychol.* 59, 14–19. doi: 10.1037/0003-066X.59.1.14

Mulder, Y. G., Lazonder, A. W., and de Jong, T. (2010). Finding out how they find it out: an empirical analysis of inquiry learners' need for support. *Int. J. Sci. Educ.* 32, 2033–2053. doi: 10.1080/09500690903289993

Narciss, S. (2006). "Informatives tutorielles Feedback," in *Entwicklungs-Und Evaluationsprinzipien Auf Der Basis Instruktionspsychologischer Erkenntnisse,* Münster:Waxmann.

Narciss, S. (2008). "Feedback strategies for interactive learning tasks," in *Handbook of Research on Educational Communications and Technology*, 3rd Edn., eds J. M. Spector, M. D. Merrill, J. van Merrienboer, and M. P. Driscoll (New York, NY: Lawrence Erlbaum Associates), 125–144. doi: 10.4324/9780203880869.ch11

Nye, B. D., Graesser, A. C., and Hu, X. (2014). "Multimedia learning with intelligent tutoring systems," in *The Cambridge Handbook of Multimedia Learning*, ed R. Mayer (Cambridge: Cambridge University Press), 705–728. doi: 10.1017/CBO9781139547369.035

Pedaste, M., Lazonder, A. W., Raes, A., Wajeman, C., Moore, E., and Girault, I. (2015). "Grand challenge problem 3: empowering science teachers using technology-enhanced scaffolding to improve inquiry learning," in *Grand Challenge Problems in Technology Enhanced Learning II: MOOCS and Beyond: Perspectives for Research, Practice, and Policy Making Developed at the Alpine Rendez- Vous in Villard-de-Lans,* eds K. Lund, P. Tchounikine, and F. Fischer (Cham: Springer Briefs in Education), 18–20.

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., et al. (2015). Phases of inquiry-based learning: definitions and the inquiry cycle. *Educ. Res. Rev.* 14, 47–61. doi: 10.1016/j.edurev.2015.02.003

Prince, M. J., and Felder, R. M. (2007). The many faces of inductive teaching and learning. *J. Coll. Sci. Teach.* 36, 14–20. Available online at: www.jstor.org/stable/42992681

Quinn, M. E., and George, K. D. (1975). Teaching hypothesis formation. *Sci. Educ.* 59, 289–296. doi: 10.1002/sce.3730590303

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna. Available online at: https://www.r-project.org/

Ryoo, K., and Linn, M. C. (2016). Designing automated guidance for concept diagrams in inquiry instruction. *J. Res. Sci. Teach*. 53, 1003–1035. doi: 10.1002/tea.21321

Shermis, M. D., and Burstein, J. (2013). *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. New York, NY: Routledge. doi: 10.4324/9780203122761

Shieber, S. M. (2003). *An Introduction to Unification-Based Approaches to Grammar*. Brookline, MA: Microtome Publishing.

Timmers, C. F., Walraven, A., and Veldkamp, B. P. (2015). The effect of regulation feedback in a computer-based formative assessment on information problem solving. *Comp. Educ.* 87, 1–9. doi: 10.1016/j.compedu.2015.03.012

Tschirgi, J. E. (1980). Sensible reasoning: a hypothesis about hypotheses. *Child Dev.* 51, 1–10. doi: 10.2307/1129583

Van der Kleij, F. M., Feskens, R. C. W., and Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: a meta-analysis. *Rev. Educ. Res.* 85, 475–511. doi: 10.3102/0034654314564881

Van Joolingen, W. R., and De Jong, T. (1991). Supporting hypothesis generation by learners exploring an interactive computer simulation. *Instruct. Sci.* 20, 389–404. doi: 10.1007/BF00116355

Van Joolingen, W. R., and De Jong, T. (1993). Exploring a domain with a computer simulation: traversing variable and relation space with the help of a hypothesis scratchpad. *Simulat. Based Exp. Learn.* 191–206. doi: 10.1007/978-3-642-78539-9_14

van Riesen, S. A. N., Gijlers, H., Anjewierden, A., and de Jong, T. (2018). The influence of prior knowledge on experiment design guidance in a science inquiry context. *Int. J. Sci. Educ.* 40, 1327–1344. doi: 10.1080/09500693.2018.1477263

Vandewaetere, M., Desmet, P., and Clarebout, G. (2011). The contribution of learner characteristics in the development of computer-based adaptive learning environments. *Comput. Human Behav*. 27, 118–130. doi: 10.1016/j.chb.2010.07.038

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Dev. Rev.* 27, 172–223. doi: 10.1016/j.dr.2006.12.001

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Education

ORIGINAL RESEARCH
published: 04 April 2019
doi: 10.3389/feduc.2019.00028

The Influence of Variance in Learner Answers on Automatic Content Scoring

Andrea Horbach and Torsten Zesch*

Language Technology Lab, University Duisburg-Essen, Duisburg, Germany

Automatic content scoring is an important application in the area of automatic educational assessment. Short texts written by learners are scored based on their content while spelling and grammar mistakes are usually ignored. The difficulty of automatically scoring such texts varies according to the variance within the learner answers. In this paper, we first discuss factors that influence variance in learner answers, so that practitioners can better estimate if automatic scoring might be applicable to their usage scenario. We then compare the two main paradigms in content scoring: (i) similarity-based and (ii) instance-based methods, and discuss how well they can deal with each of the variance-inducing factors described before.

Keywords: automatic content scoring, short-answer questions, natural language processing, linguistic variance, machine learning

OPEN ACCESS

Edited by:
Ronny Scherer,
Department of Teacher Education and School Research, Faculty of Educational Sciences, University of Oslo, Norway

Reviewed by:
Mark Gierl,
University of Alberta, Canada
Dirk Ifenthaler,
Universität Mannheim, Germany

****Correspondence:***
Andrea Horbach
andrea.horbach@uni-due.de

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Education

Received: *07 December 2018*
Accepted: *12 March 2019*
Published: *04 April 2019*

Citation:
Horbach A and Zesch T (2019) The Influence of Variance in Learner Answers on Automatic Content Scoring. Front. Educ. 4:28. doi: 10.3389/feduc.2019.00028

1. INTRODUCTION

Automatic content scoring is a task from the field of educational natural language processing (NLP). In this task, a free-text answer written by students should be automatically assigned a score or correctness label in the same way as a human teacher would do. Content scoring tasks have been a popular exercise type for a variety of subjects and educational scenarios, such as listening or reading comprehension (in language learning) or definition questions (in science education). In a traditional classroom-setting, answers to such exercises are manually scored by a teacher, but in recent years, their automatic scoring has received growing attention as well (for an overview, see e.g., Ziai et al., 2012 and Burrows et al. (2014)). Automatic content scoring may decrease the manual scoring workload (Burstein et al., 2001) as well as offer more consistency in scoring (Haley et al., 2007). Additionally, automatic scoring provides the advantage that evaluation can happen in the absence of a teacher so that students may receive feedback immediately without having to wait for human scoring. With the increasing popularity of MOOCS and other online learning platforms, automatic scoring has become a topic of growing importance for educators in general.

In this paper, we restrict ourselves to short-answer questions as one instance of free-form assessment. While other test types, such as multiple choice items, are much easier to score, free-text items have an advantage from a testing perspective. They require active formulation instead of just selecting the correct answer from a set of alternatives, i.e., they test production instead of recognition.

Answers to short-answer questions have a typical length between a single phrase and two to three sentences. This places them in length between gap-filling exercises, which often ask for single words, and essays, where learners write longer texts. We do not cover automatic essay scoring in this article, even if it is related to short-answer scoring, and to some extent even the same methods might be applied. The main reason is that scoring essays usually takes into consideration the form of the essay (style, grammar, spelling, etc.) in addition to content (Burstein et al., 2013), which introduces many additional factors of influence that are beyond our scope.

Figure 1 shows examples from three different content scoring datasets (ASAP, POWERGRADING and SEMEVAL) and highlights the main components of a content scoring scenario: a prompt, a set of learner answers with scoring labels, and (one or several) reference answers.

- A **prompt** consists of a particular question and optionally some textual or graphical material the question is about (this additional material is omitted in **Table 1** for space reasons).
- A set of **learner answers** that are given in response to that prompt. The learner answers in our example have different length ranging from short phrasal answers in POWERGRADING to short paragraphs in ASAP. They may also contain spelling or grammatical errors. As discussed above, these errors should not be taken into consideration when scoring an answer.
- The task of automatic scoring is to assign a **scoring label** to a learner answer. If we want to learn such an assignment mechanism, we typically need some scored examples, i.e., learner answers with a gold-standard scoring label assigned by a human. As we can see in the example, the kind of label varies between datasets and can be either numeric or categorial, depending on the nature of the task and also of the purpose of the automatic scoring.

  Numeric or binary scoring labels, as we see in ASAP and POWERGRADING, can be easily summed up and compared. They are thus often used in summative feedback, where the goal is to inform teachers, e.g., about the performance of students in a homework assignment. For formative feedback, which is directed toward the learner, in contrast, a more informative categorical label might be preferable, e.g., to inform a student of their learning progress. The SEMEVAL data is an example for scoring labels aiming into that direction.
- In addition to learner answers, datasets often include teacher-specified **reference answers** for each label. A reference answer showcases a representative answer for a given score and can be used for (human or automatic) comparison with a learner answer. Alternatively, scoring guidelines describing properties of answers with a certain score can be provided. This is often the case when answers are so complex that just providing a small number of reference answers does not nearly cover the conceptual range of possible correct answers and misconceptions. This is for example the case for the ASAP dataset. When reference answers are given, many datasets only provide reference answers for correct answers and not for incorrect ones, e.g., POWERGRADING and SEMEVAL.

The content scoring scenario with its interrelated textual components – a prompt, learner answers, and a reference answer – render automatic content scoring a challenging application of Natural Language Processing which bears strong resemblances to various core NLP fields like paraphrasing (Bhagat and Hovy, 2013), textual entailment (Dagan et al., 2013), and textual similarity (Bär et al., 2012). In all those fields, the semantic relation between two texts is assessed, a method that directly transfers to the comparison between learner and reference answers, as we will see later.

During recent years, many approaches for automatic content scoring have been published on various datasets (see Burrows et al. (2014) for an overview). A practitioner who is considering using automatic scoring for their own educational data might easily feel overwhelmed. They might find it hard to compare approaches and draw conclusions for their applicability on their specific scoring scenario. In particular, approaches often apply various machine learning methods with a variety of features and are trained and evaluated using different datasets. Thus, comparing any two approaches from the literature can be difficult.

This paper aims to shed light on the individual factors influencing automatic content scoring and identifies the variance in the answers as one key factor that makes scoring difficult. We start in section 2 by discussing the nature of this variance, followed by a discussion of datasets and their parameters that influence variance. We discuss in section 3 properties of automatic scoring methods and review existing approaches, especially with respect to whether they score answers based on features extracted from the answers themselves or based on a comparison with a reference answer. We then discuss in section 4 how these factors can be isolated in scoring experiments. We either provide own experiments, discuss relevant studies from the literature, or formulate requirements for datasets that would make currently infeasible experiments possible.

2. VARIANCE IN LEARNER ANSWERS

Variance is the reasons why automatic scoring has to go beyond simply matching learner answers to reference answers. The more variance we find in the learner answers, the more complex the scoring model has to be and therefore the harder is the content scoring task (Padó, 2016). Thus, in this section, we discuss *why* variance increases the difficulty of automatic scoring and analyze publicly available datasets with respect to the variance-inducing properties.

2.1. Sources of Variance

From an NLP perspective, automating content scoring of free-text prompts is a challenging task, mainly due to the textual variance of answers given by the learner. Variance can occur on several levels, as highlighted in **Figure 2**. It can occur both on the conceptual level as well as on the realization level, whereas variance in realization can mean variance of the linguistic expression as well as orthographic variance.

2.1.1. Conceptual Variance

Conceptual variance occurs when a prompt asks for multiple aspects or has more than one correct solution. For example, in the prompt *Name one state that borders Mexico* from the POWERGRADING dataset, there are four different correct solutions: *California*, *Arizona*, *New Mexico*, and *Texas*. A scoring method needs to take all of them into account. However, conceptually different correct solutions are not the main problem, as their number is usually rather small. The

POWERGRADING DATASET – PROMPT 4

QUESTION: *What is the economic system in the United States?*

REFERENCE ANSWERS:
- **R**$_1$: *capitalist economy*
- **R**$_2$: *market economy*

LEARNER ANSWERS:
- **L**$_1$: *free market* **correct**
- **L**$_2$: *capitalism* **correct**
- **L**$_3$: *democratic* **correct**
- **L**$_4$: *the federal currency system* **incorrect**
- **L**$_5$: *a bad one* **incorrect**

SEMEVAL DATASET – PROMPT "VOLTAGE_DEFINE_Q"

QUESTION: *What is voltage?*

REFERENCE ANSWERS:
- **R**$_1$: *Voltage is the difference in electrical states between two terminals*

LEARNER ANSWERS:
- **L**$_1$: *is the difference in electrial stat between terminals* **correct**
- **L**$_2$: *the is a difference in the terminals* **partially_correct_incomplete**
- **L**$_3$: *the measurment of power to a source of energy* **contradictory**

ASAP DATASET – PROMPT 1

QUESTION: *After reading the group's procedure, describe what additional information you would need in order to replicate the experiment. Make sure to include at least three pieces of information.*

LEARNER ANSWERS:
- **L**$_1$: *Some additional information you will need are the material. You also need to know the size of the contaneir to measure how the acid rain effected it. You need to know how much vineager is used for each sample. Another thing that would help is to know how big the sample stones are by measureing the best possible way.* **3 points**
- **L**$_2$: *After reading the expirement, I realized that the additional information you need to replicate the expireiment is one, the amant of vinegar you poured in each container, two, label the containers before you start yar expirement and three, write a conclusion to make sure yar results are accurate.* **1 point**
- **L**$_3$: *The student should list what rock is better and what rock is the worse in the procedure.* **0 points**

FIGURE 1 | Exemplary content scoring prompts from three different datasets with reference answers (if available) as well as several learner answers with their scoring labels.

much bigger problem is variance within incorrect answers, as there are usually many ways for a learner to get an answer wrong so that incorrect answers often correspond to several misconceptions. For the Powergrading example prompt in **Table 1** (asking *What is the economic system in the United States?*), frequent misconceptions center around *democracy* or *US dollar*, but there also is a long tail of infrequent other misconceptions.

2.1.2. Variance in Realization

In contrast to the conceptual variance we have just discussed, which covers different ways of conceptually answering a question, variance in realization means different ways of formulating the same conceptual answer. We consider variance in linguistic expression as well as variance on the orthographic level.

2.1.2.1. Variance of linguistic expression

This refers to the fact that natural language provides many possibilities to express roughly the same meaning (Meecham and Rees-Miller, 2005; Bhagat and Hovy, 2013). This variance of expression makes it in most cases impossible to preemptively enumerate all correct solutions to a prompt and score new learner answers by string comparison alone. For example consider the

TABLE 1 | Dataset statistics.

Corpus	# Answers	# Prompts	∅ tokens/answer		
			min	med	max
ASAP	33,320	10	26.5	48.5	66.2
ASAP-DE	903	3	24.6	33.0	33.9
CREE	566	62	5.7	21.6	68.1
CREG	1,032	177	5.0	9.7	45.8
CS	630	21	6.2	20.6	36.0
CSSAG	1,840	31	10.9	23.5	42.6
Powergrading	6,980	10	1.9	3.4	8.4
PT_ASAG	3,675	15	9.5	14.3	40.8
SRA	5,239	182	3.4	11.7	44.3

Tokens per answer are counted individually across all answers for one prompt and the minimum, median[1], and maximum of these values reported. i.e., the prompt with the shortest answers in ASAP has on average 26.5 tokens.

following three sentences. They all come from the SemEval prompt in **Figure 1**. The first is a reference answer, while the other two are learner answers.

- **R** *Voltage is the difference in electrical states between two terminals*
- L_1 *[Voltage] is the difference in electrial stat between terminals*
- L_2 *[Voltage is] the measurement between the electrical states of the positive and negative terminals of a battery.*

While the first learner answer in the example above shares many words with the reference answer, the second learner answer has much lower overlap. The term *difference* is replaced by the related term *measurement*. For such cases of lexical variance, we need some form of external knowledge to decide that *difference* and *measurement* are similar.

2.1.2.2. Orthographic variance

A property of (especially non-native) learner data that also contributes toward high realization variance in the data is the orthographic variability and occurrence of linguistic deviations from the standard (Ellis and Barkhuizen, 2005), which can also make it hard for humans to understand what was intended (Reznicek et al., 2013). For example in the learner answer L_1 above, the learner misspelled *electrical state* as *electrial stat*. The number of spelling errors – and thus how pronounced this deviation is – depends on a number of factors, such as whether answers have been written by language learners or native speakers or whether answers refer to a text visually available to the learner at the time of writing the answer or not.

2.2. Content Scoring Datasets

In the following, we introduce publicly available datasets for content scoring. Afterwards, we categorize all datasets in **Tables 1**, **2** according to various factors that influence variance. The datasets come from different research contexts, we present them here in alphabetical order:

- The **Asap dataset**[2] has been released for the purpose of a scoring competition and contains answers collected at US high schools for 10 different ppts from various subjects. The main distinguishing features for this dataset are the large number of answers per individual prompt as well as relative high length of answers. A German version of the dataset, **Asap-De**, addressing three of the science prompts, has been collected by Horbach et al. (2018).
- The **CREE dataset** (Bailey and Meurers, 2008) contains answers given by learners of English as a foreign language for reading comprehension questions. The number of answers per prompt as well as the overall number of learner answers in this dataset is comparably low.
- The **CREG dataset** (Meurers et al., 2011a) is similar to CREE in that it targets reading comprehension questions for foreign language learners, but here the data is in German, so it is an instance of a non-English dataset. Answers were given by beginning and intermediate German-as-a-foreign-language learners at two US universities and respond to reading comprehension questions.
- The **CS dataset** (Mohler and Mihalcea, 2009) contains answers to computer science questions given by participants of a university course. In this dataset, the questions stand alone and do not address additional material, such as reading texts or experiment descriptions.
- The **CSSAG dataset** (Pado and Kiefer, 2015) contains computer science questions collected from participants of a university-level computer-science class in German.
- The **Powergrading dataset** (Basu et al., 2013) addresses questions from US immigration exams and learner answers have been crowd-sourced. It is unclear what the language proficiency of the writers is, including whether they are native speakers or not. The dataset contains the shortest learner answers of all datasets.
- The Portuguese **PT_ASAG dataset** (Galhardi et al., 2018) contains learner answers collected in biology classes at schools in Brazil using a web system. Apart from reference answers for each question, the dataset also contains keywords specifying aspects of a good question.
- The **Student Response Analysis (SRA) dataset** (Dzikovska et al., 2013) was used in SemEval-2013 shared task. It consists of two subsets, both dealing with science questions: The Beetle subset covers student interactions with a tutoring system, while the SciEntsBank subset contains answers to assessment questions. A special feature of this dataset is that learner answers are annotated with three different types of labels: (i) binary correct/incorrect decisions, (ii) with categories used for recognizing textual entailment such as whether an answer entails or contradicts the reference answer), as well as (iii) formative assessment labels, informing students, e.g., that an answer is partially correct, but incomplete.

[1] For the median, we report the lower median if there is an even number of items, so that the value corresponds to the average number of tokens per answer of a specific prompt.

[2] https://www.kaggle.com/c/asap-sas

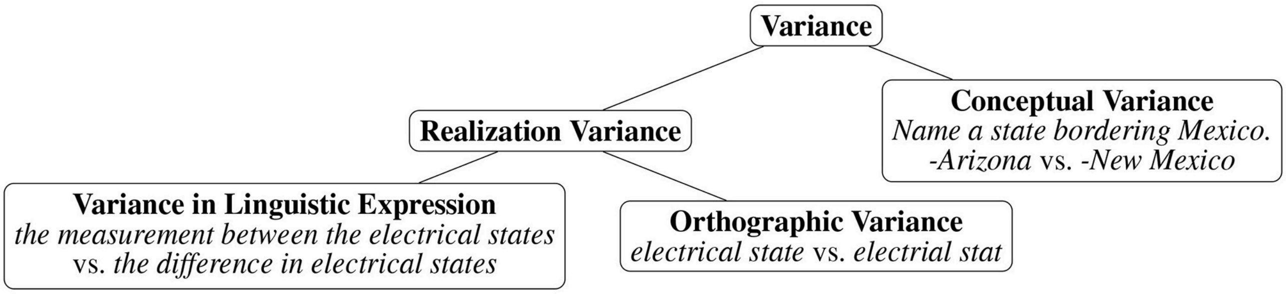

Variance
Realization Variance
Conceptual Variance
Name a state bordering Mexico.
-Arizona vs. -New Mexico
Variance in Linguistic Expression
the measurement between the electrical states
vs. the difference in electrical states
Orthographic Variance
electrical state vs. electrial stat

FIGURE 2 | Sources of variance in content scoring.

TABLE 2 | Overview of content scoring datasets.

Corpus	Prompt type	Language	Learner population	Scoring labels
ASAP	Sciences, biology, reading comprehension	English	High school students	Numeric [0, 1, 2, (3)]
ASAP-DE	Sciences	German	Crowdworkers	Numeric [0, 1, 2, (3)]
CREE	Reading comprehension for language learning	English	university students learning English	Binary & diagnostic
CREG	Reading comprehension for language learning	German	US university students learning German	Binary & diagnostic
CS	Computer science questions	English	university students	Numeric [0, 0.5, . . . , 5]
CSSAG	Computer science	German	University students	Numeric [0, 0.5, . . . , 2]
Powergrading	Immigration exams	English	Unknown (crowdworkers)	Binary
PT_ASAG	Biology	Portuguese	8th & 9th grade students	Numeric [0, 1, 2, (3)]
SRA	Science questions	English	High school students	Entailment labels (binary & diagnostic)

2.3. Dataset Properties Influencing Variance

We now discuss dataset-inherent properties that can help us to estimate the amount of variance to be expected in data.

2.3.1. Prompt Type

The type of prompt has a strong influence on the expected answer variance. Imagine, for example, a factual question like *Where was Mozart born?* and a reading comprehension question such as *What conclusion can you draw from the text?* For the first question, there is no variance in the correct answers (*Salzburg*) and probably only little variance in the misconceptions (*Vienna*). For the second question, a very high variance is to be expected. In general, the more open-ended a question is, the harder it will be to automatize its scoring.

Different answer taxonomies have been proposed to classify questions in the classroom according to the cognitive processes involved for the student and they provide also clues about ease of automatic scoring. Anderson et al. (2001) provide a classification scheme according to the cognitive skills that are involved in solving an exercise: remembering, understanding, applying, analyzing, evaluating, and creating in ascending order of difficulty for the student. This taxonomy could of course also be applied to content scoring prompts. Padó (2017) annotates questions in the CSSAG dataset according to this taxonomy and finds that questions from the lower categories are not only easier for students, but produce also less variance and need less elaborate methods for automatic scoring. She also finds that the instructional context of a question needs to be considered when assigning a level (e.g., to differentiate between a real analyzing question and one that is actually a remembering question because the analysis has been explicitly made in the course). Therefore it is hard to apply such a taxonomy to a dataset where the instructional context is unknown.

A taxonomy specifically for reading comprehension questions has been developed by Day and Park (2005). It classifies questions by comprehension as literal, reorganization, inference, prediction, evaluation, and personal response (again ordered from easy to hard). Literal questions are the easiest because their answers can be found verbatim in the text. Such questions tend to have lower variance, especially when given to low-proficiency learners, as they often lift their answers from the text. Also for this taxonomy, it has been found that reading comprehension prompts for language learners focus on the lower comprehension types (Meurers et al., 2011b) and that among these literal questions are easier to score than reorganization and inference questions. We argue that questions with comprehension types higher in the taxonomy contain so much variance that they are difficult to handle automatically. An example for a personal response question from Day and Park (2005) is *What do you like or dislike about this article?* We argue that answers to such questions go beyond content-based evaluation and rather touch the area of essay scoring, as how an opinion is expressed it might be more important than its actual content.

The modality of a prompt also plays a role. By modality, we mean whether a question refers to a written or a spoken text. Especially for non-native speakers, listening comprehension exercises will yield a much higher variance as learners cannot copy material from the text based on the written form, but mostly write what they think they understood auditorily. This leads especially to a high orthographic variance and makes scoring harder compared to a similar prompt administered as reading comprehension exercise.

Table 2 shows that existing datasets cover very diverse prompts from reading comprehension for language learning over science question to biology and literature questions, but that they do not nearly cover all possible prompt types.

2.3.2. Answer Length

Answer length of course is strongly related to the type of question asked. *Where* or *when* questions usually require only a phrasal answer, whereas *why* questions are often answered with complete sentences. Shorter answers consisting of only a few words often correspond only to a single concept mentioned in the answer (see the example from the POWERGRADING dataset in **Table 1**), whereas longer answers (as we saw in the ASAP example) tend to be also conceptually more complex. It seems intuitive that this conceptual complexity is accompanied by a higher variance in the data. In a longer answer, there are more options how to phrase and order ideas in different ways.

Answer length is a measure that can be easily determined for a new dataset once the learner answers are collected, so it can serve as a quick indicator for the ease of scoring. In general, shorter answers can be scored better than longer answers. Of course, also datasets with answers of the same length can display different types of complexity and variance. Nevertheless, we consider answer length as a good and at the same time cheap indicator.

Table 1 presents some core answer length statistics for each dataset. A dataset usually consists of several individual prompts and different prompts in a dataset might differ more or less from each other. To characterize the variance between prompts in a dataset better we give the average answer length in tokens, as well as the minimum, median, and maximum value across the different prompts. **Figure 3** visualizes for each dataset the distribution of the average answer length per prompt. We see that the individual datasets span a wide range of lengths from very short phrasal answers in POWERGRADING to long answers almost resembling short essays in ASAP. We also see that the number of different prompts and individual learner answers and thus also the number of learner answers for each prompt varies considerably, from datasets with only a very restricted number of answers for each question, such as in CREE and CREG, to several thousand answers per prompt in ASAP.

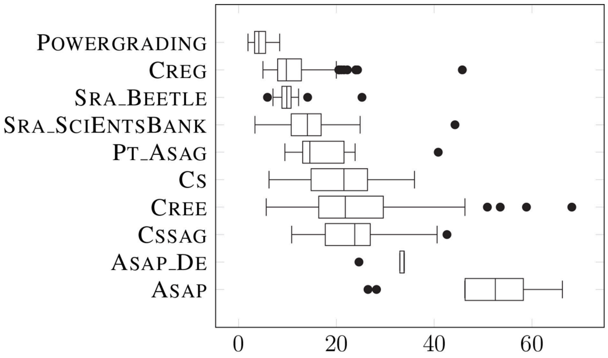


FIGURE 3 | Average lengths of answers per dataset.

2.3.3. Language

The language that is used to answer a prompt, such as English, German, or Chinese, is also an important factor influencing the answer variance. Methods that work well for one language may not be directly transferable to other languages. This is due both to the linguistic properties of individual languages as well as to the availability of language-specific NLP resources used for scoring. By linguistic properties we mean especially the morphological richness of a language and the restrictiveness of word order. If an answer given in English talks about a red apple, it might be sufficient to look for the term *red apple*, while in German, depending on the grammatical context, terms such as *(ein) roter Apfel*, *(der) rote Apfel*, *(einen) roten Apfel*, or *(des) roten Apfels* might occur. Thus, a scoring approach based on token n-grams usually needs fewer training instances in English compared to German, as an English n-gram often corresponds to several German n-grams. For morphologically-richer languages such as Finnish or Turkish, approaches developed for English might completely fail.

Freeness of word order is related to morphological richness. Highly inflected languages, such as German, have usually a less restricted word order than English. Thus, n-gram models work well for the mainly linear grammatical structures in English, but less so for German with freer word-order and more long-distance dependencies (Andresen and Zinsmeister, 2017).

As for language resources used in content scoring methods, there are two main areas which have to be considered: linguistic processing tools as well as external resources. Many scoring methods rely on some sort of linguistic processing. The automatic detection of word and sentence boundaries (tokenization) is a minimal requirement necessary for almost all approaches, while some methods additionally use for example lemmatization (detecting the base form of a word), part-of-speech-tagging (labeling words as nouns, verbs or adjectives), or parsing sentences into syntax trees, which represent the internal linguistic structure of a sentence. External resources can be, for example, dictionaries used for spellchecking, but also resources providing information about the similarity between words in a language.

Coming back to the example above, to know that *measurement* and *difference* are related, one would either need an ontology crafted by an expert, such as WordNet (Fellbaum, 1998), or would need similarity information derived from large corpora, based on the core observation in distributional semantics that words are similar if they often appear in similar contexts (Firth, 1957). The availability of such tools and resources has to be taken into consideration when planning automatic scoring for a new language.

2.3.4. Learner Population and Language Proficiency

The learner population is another important factor to consider, as it defines the language proficiency of the learners, i.e., whether they are beginning foreign language learners or highly proficient native speakers. Language proficiency can have two, at first glance contradicting, effects: A low language proficiency might lead to a high variance in terms of orthography, because beginners are more likely to make spelling or grammatical errors. At the same time, being a low-proficiency learner, can equally reduce variance, but on the lexical and syntactic level. This is because such a learner will have a more restricted vocabulary and has acquired fewer grammatical constructions than a native speaker. Moreover, low-proficiency learners might stay closer to the formulations in the prompt, especially when dealing with reading comprehension exercises, where the process of re-using material from the text for an answer is known as "lifting."

Beginning language learners and fully proficient students are of course only the far points of the scale, while students from different grades in school would rank somewhere in between. **Table 2** shows that the discussed datasets indeed cover a wide range of language proficiencies.

Also the homogeneity of the learner population plays a role: Learners from a homogeneous population can be expected to produce more homogeneous answers. It has, for example, been shown that the native language of a language learner influences the errors a learner makes (Ringbom and Jarvis, 2009). A German learner of English might be more inclined to misspell the word *marmalade* as *marmelade* because of the German cognate *Marmelade* (Beinborn et al., 2016). An automatic scoring engine trained on learner answers given by German learners might thus encounter the misspelling *marmelade* often enough to learn that an answer containing this word is as good as an answer containing the right spelling. However, a model trained on answers by learners from many different countries might not be able to learn (partially overlapping) error patterns for each individual first language of the learners. In a slightly different way, this also applies to native speakers. Consider e.g., answers by students from one university which all attended the same lecture and used the same slides and textbooks for studying (low variance) vs. answers by students from different universities using different learning materials (high variance).

2.3.5. Other Factors

The following factors do not directly influence the variance found in the data, but are other data-inherent factors that influence the difficulty of automatic scoring.

2.3.5.1. Dataset size

When using machine learning models to perform content scoring, as do all the approaches we discuss in this article, the availability of already-scored answers from which the scoring method can learn is an important parameter (Heilman and Madnani, 2015). The more answers there are to learn from, the better we can usually model what a correct or incorrect answer looks like. The range of available answers covered varies between less than 10 answers for a prompt (as for example in the CREG dataset where a model across individual questions is learnt by most approaches dealing with this dataset) and over 3,000 answers per prompt in the ASAP dataset.

In many practical settings, only a small part of the available data is manually scored and used for training. It has been shown that the choice of training data heavily influences scoring performance and that the variance within the instances selected for training is a major influencing factor (Zesch et al., 2015a; Horbach and Palmer, 2016).

2.3.5.2. Label set

Different label sets have been proposed for different content scoring datasets. The educational purpose of the scoring scenario is the main determining factor for this choice. Some datasets such as CREG and SRA have even more than one label set so that different usage scenarios can be addressed. This purpose can either be to generate *summative* or *formative feedback* (Scriven, 1967). The recipient of summative feedback is the teacher who wants to get an overview of the performance of a number of learners, for example in a placement test or exam situation. In this case, it is important that scores are comparable and can be aggregated so that there is an overall result for a test consisting of several prompts. Binary or numeric scores fit this purpose well. Formative feedback in contrast, as given through the categorical labels in SRA, CREG, and CREE, is directed toward the learner and meant to inform learners about their progress and the problems they might have had with answering a question. This type of feedback in content scoring is, for example, used in automatic tutoring systems. For a learner, the information that she scored 3.5 out of 5 points might be not as informative as a more meaningful feedback message stating that she missed an important concept required in a correct answer. Thus, datasets meant for formative feedback often use categorical labels rather than numeric ones.

The kind of label that is to be predicted obviously influences the scoring difficulty. In general, the more fine-grained the labels, the harder they are to predict given the same overall amount of training data. Also the conceptual spread covered by the labels can make the task more or less difficult. If the labels intend to make very subtle distinctions between similar concepts, the task is more complex than a scoring scheme that differentiates between coarser categories and considers everything as correct that is somewhat related to the correct answer.

2.3.5.3. Difficulty of the scoring task for humans

All machine learning algorithms learn from a gold-standard produced by having human experts (such as teachers) label the data. If the scoring task is difficult, humans will make errors and label data inconsistently. This noise in the data impedes performance of a machine learning algorithm. If the gold-standard dataset is constructed from two trained human annotators, the inter-annotator-agreement between these two is considered to be an upper bound of the performance that can be expected from a machine. If two teachers agree only in 90% of the scores they assign for the same task, 90% agreement with the gold-standard is also considered the best possible result obtainable by automatic scoring (Gale et al., 1992; Resnik and Lin, 2010). The same argument can be applied for self-consistency. If a teacher labels the same data twice and can reproduce his own cores only for 90% of all answers, we can consider this 90% an upper bound for machine learning. This influence parameter obviously depends on most of the others and cannot be considered in isolation, but it helps to estimate which level of performance is to be expected for a particular prompt.

2.4. Summary

In this section, we have discussed several factors that are influencing the variance to be found in learner answers: the prompt type, answer length, language and learner population. We also introduced dataset size, the label set and the scoring difficulty for human scorers as additional parameters that influence the suitability of a dataset for human scoring. In the next section, we first give an overview of content scoring methods and then present a set of experiments that show the influence of some of the discussed factors on content scoring.

3. AUTOMATIC CONTENT SCORING

As explained in the introduction, the overall aim of content scoring is to mimic a teacher's scoring behavior by assigning labels to a learner answers indicating how good the answer is content-wise.

A very large number of automatic content scoring methods have been proposed (see Burrows et al., 2014 for an overview), but we argue that most existing methods can be categorized into two main paradigms: similarity-based and instance-based scoring. Hence, instead of analyzing the properties of single scoring methods, we can draw interesting conclusions by comparing the two paradigms.

3.1. Similarity-Based Approaches

Figure 4 gives a schematic overview of similarity-based scoring. The learner answer is compared with a reference answer (or a high-scoring learner answer) based on a similarity metric. If the similarity surpasses a certain threshold (exemplified by 0.7 in **Figure 4**), the learner answer is considered as correct. Note that reference answers are always examples for correct answers. In the datasets discussed in section 2.2, there are no samples for incorrect answers, although we have seen earlier that also incorrect answers might form groups of answers expressing the same content.

An important factor in the performance of such similarity-based approaches is how the similarity between answers is computed. In the simplest form, it can be computed based on surface overlap, such as token overlap, where the amount of words or characters shared between answers is measured or edit distance, where the number of editing steps necessary to transform one answer into another is counted. These methods work well when different correct answers can be expected to mainly employ the same lexical material. However, when paraphrases are expected to be lexically diverse, surface-based methods might not be optimal. Consider the hypothetical sentence pair *Paul presented his mother with a book - Mary received a novel from her son as a gift*. In such a case the overlap between the two sentences on the surface is low, while it is clear to human readers that the two sentences convey a very similar meaning. To retrieve the information that *present* and *gift* from the above example are highly similar, semantic similarity methods make use of ontologies like WordNet Fellbaum (1998) or large background corpora [e.g., latent semantic analysis (Landauer and Dumais, 1997)].

In the content scoring literature, all these kinds of similarities are used. While Meurers et al. (2011c) mainly rely on similarity on the surface level for different linguistic units (tokens, chunks, dependency triples), methods such as Mohler and Mihalcea (2009) rely on external knowledge about semantic similarity between words.

3.2. Instance-Based Approaches

In instance-based approaches, lexical properties of correct answers (words, phrases, or even parts of words) are learned from other learner answers labeled as correct, while commonalities between incorrect answers inform the classifier about common misconceptions in learner answers. One would, for example, as depicted in **Figure 5**, learn that certain n-grams, such as *electrical states*, are indicators for correct answers while others, such as *battery*, are indicators for incorrect answers. For the scoring process, learner answers are then represented as feature vectors where each feature represents the occurrence of one such n-gram. The information about good n-grams is prompt-specific. For a different prompt, such as one asking for the power source in a certain experiment, *battery* might indicate a good answer, while answers containing the bigram *electrical states* would likely be wrong.

As the knowledge used for classification usually comes from the dataset itself and, in many approaches, no external knowledge is used in the scoring process (in contrast to similarity-based scoring), instance-based methods tend to need more training data and do not generalize as well across prompts. Instance-based methods have been used, for example, for various work on the ASAP dataset (Higgins et al., 2014; Zesch et al., 2015b), including all the top-performing systems from the ASAP scoring competition (Conort, 2012; Jesensky, 2012; Tandalla, 2012; Zbontar, 2012), as well as in commercially used systems.

3.3. Comparison

We presented two conceptually different ways of content scoring, one relying on the similarity with a reference answer (similarity-based) and the other on information about lexical material in the learner answers (instance-based). While we have presented the

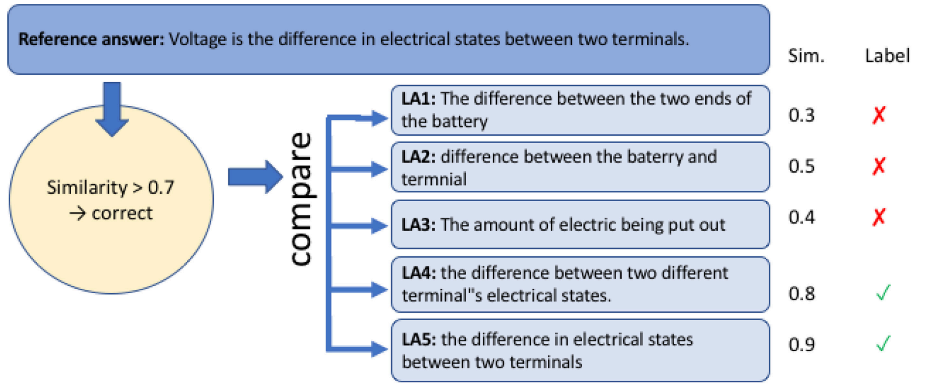

Reference answer: Voltage is the difference in electrical states between two terminals.
Sim.
Label
Similarity > 0.7
→ correct
compare
LA1: The difference between the two ends of the battery
0.3
✗
LA2: difference between the baterry and termnial
0.5
✗
LA3: The amount of electric being put out
0.4
✗
LA4: the difference between two different terminal"s electrical states.
0.8
✓
LA5: the difference in electrical states between two terminals
0.9
✓

FIGURE 4 | Schematic overview of similarity-based scoring.

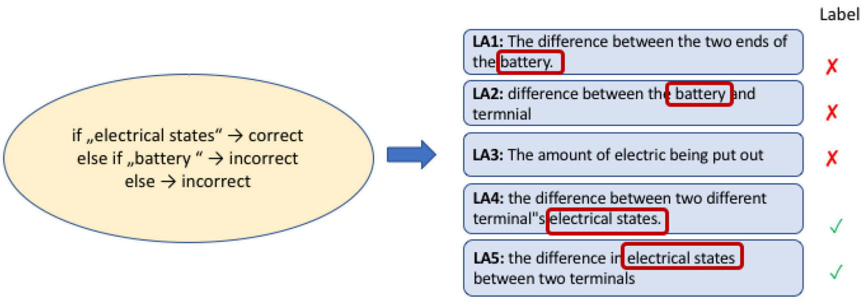

Label
if „electrical states" → correct
else if „battery " → incorrect
else → incorrect
LA1: The difference between the two ends of the battery.
✗
LA2: difference between the battery and termnial
✗
LA3: The amount of electric being put out
✗
LA4: the difference between two different terminal"s electrical states.
✓
LA5: the difference in electrical states between two terminals
✓

FIGURE 5 | Schematic overview of instance-based scoring.

paradigmatic case for each side, there are of course less clear-cut cases. For example, an instance-based k-nearest-neighbor classifier scores new unlabeled answers by assigning them the label of the closest labeled learner answer. By doing so the classier inherently exploits similarities between answers.

3.3.1. Associated Machine Learning Approaches

Classical supervised machine learning approaches have been associated with both types of scoring paradigms. Instance-based approaches often work on feature vectors representing lexical items, while similarity-based approaches (Meurers et al., 2011c; Mohler et al., 2011) use various overlap measures as features or rely on just one similarity metric (Mohler and Mihalcea, 2009). Deep learning methods have been applied for instance-based scoring Riordan et al. (2017) as well as similarity-based scoring Patil and Agrawal (2018). As content scoring datasets are often rather small, the performance gain by using deep learning methods has far not been as in other NLP areas, if there was a reported gain at all.

3.3.2. Source of Knowledge

In general, instance-based approaches mainly use lexical material present in the answers while similarity-based methods often leverage external knowledge resources like WordNet or distributional semantics to bridge the vocabulary gap between differently phrased answers. Deep learning approaches usually also make use of external knowledge in the form of embeddings that also encode similarity between words.

3.3.3. Prompt Transfer

Another aspect to consider when comparing scoring paradigms is the transferability of models to new prompts. As similarity-methods learn about a relation between two texts rather than the occurrence of certain words or word combinations, such a model can also be transferred to new prompts for which it has not been trained. For instance-based approaches, a particular word combination indicating a good answer for one prompt might not have the same importance for another prompt. We can therefore generally expect that similarity-based models transfer more easily to new prompts.

4. EXPERIMENTS AND DISCUSSION

In the previous sections, we have introduced (i) the factors influencing the variance of learner answers and the overall difficulty of the scoring task, and (ii) the two major paradigms in automatic content scoring: similarity-based and instance-based scoring. In this section, we bring both together. In the few cases where empirical evidence already exists, we direct the reader to experiments in the literature that address these influences. We design and conduct a set of experiments to explore those sources of variance that have been experimentally examined yet. However, for some dimensions of variance we have no empirical basis as evaluation datasets are sparse and do not cover the full range of necessary properties. In these cases, we instead describe desiderata for datasets that would be needed to investigate such influences. The discussion in this section is aimed at providing guidance for matching paradigms with use-cases in order to allow a practitioner to choose a setup according to the needs of their automatic scoring scenario.

4.1. Experimental Setup

Our experiments (instance-based as well as similarity-based) build on the Escrito scoring toolkit (Zesch and Horbach, 2018) (in version 0.0.1) that is implemented based on DKPro TC (Daxenberger et al., 2014) (in version 1.0.1). For preprocessing, we use DKPro Core.[3] We apply sentence splitting, tokenization, POS-tagging and lemmatization. We did not spellcheck the data, as Horbach et al. (2017) found that the amount of spelling errors in the ASAP data did not impede scoring performance in an experimental setup similar to ours.

We use a standard machine learning setup, variants of which have been used widely. We extract token and lemma n-gram features, using uni- to trigrams for tokens and bi- to four-grams for characters. We train a support vector machine using the Weka SVM classifier with SMO optimization in its standard configuration, i.e., without standard parameter tuning.

4.1.1. Datasets

We select datasets from those discussed above (see section 2.2). The main selection criterion is, that a dataset contains a high number of learner answers per prompt, so that we can investigate the influence of training data size in prompt-specific models. To meet this criterion we use POWERGRADING, ASAP, and SEMEVAL.

4.1.2. Evaluation Metric

One common type of evaluation measure applicable for all label sets in short answer scoring is accuracy, i.e., the percentage of correctly classified items. This often goes together with a per-class evaluation of precision, recall, and F-score. Kappa values, taking into account the chance agreement between the machine learning outcome and the gold standard also are quite popular. This holds especially for Quadratically Weighted Kappa (QWK) for numeric scores, as it not only considers whether an answer is correctly classified or not, but also how far of an incorrect answer is. As QWK became a quasi-standard through its usage in the Kaggle ASAP challenge, we use it for our experiments as well.

4.1.3. Learning Curves

We listed the amount of available training data as one important influence factor for scoring performance. We can simulate datasets of different sizes by using random subsamples of a dataset. By doing this iteratively several times and for several amounts of training data, we obtain a learning curve. If a classifier learns from more data results usually improve until the learning curve approximates a flat line. When we provide learning curve experiments, we always sample 100 times for each amount of training data and average over the results.

4.2. Answer Length

As to our knowledge answer length has not been examined as an influencing factor so far, we test the hypothesis that shorter answers are easier to score, as they should have less variance in general. For this purpose, we conduct experiments with increasing amounts of training data and plot the resulting learning curves. Prompts from datasets with shorter answers should converge faster and at a higher kappa than prompts with longer answers. Note that we restrict ourselves to instance-based experiments here, as there is an insufficient number of datasets providing the necessary reference answers. However, we expect the general results to also hold similarity-based methods, as the similarity of longer answers is harder to compute than for shorter answers.

Figure 6 shows the results for instance-based scoring for a number of prompts covering a wide variety of different average lengths, selected from POWERGRADING (short answers), SRA (medium length answers), and ASAP (long answers, split in two prompts with on-average about 25 tokens per answer as well as eight prompts with more than 45 tokens per answer). We observe that (as expected) shorter prompts are easier to score, but the results between individual prompts (thin lines) within a dataset vary considerably. Thus, we also present the average over all prompts from the dataset (thick line), that clearly support the hypothesis.

These experiments also tell us something about the influence of the number of training data. An obvious finding is that more data yields, for most prompts, better results. A more interesting observation is that the curves for the SRA answers level off earlier than for the ASAP and POWERGRADING datasets. This means we could not learn much more given the current machine learning algorithm, parameter settings and feature set even if we had more training data. The ASAP and POWERGRADING curves, in contrast, are still raising: if we had more training data available, we could expect a better scoring performance.

4.3. Prompt Type

In our experiments regarding answer length, we cannot fully isolate effects originating from the length of the answers from other effects like the prompt type (as some prompts require longer answers than others) and learner population (as certain prompts are suitable only for a certain learner population). Therefore, we now try to isolate the effect of the prompt type by choosing prompts with answers of the same length and coming from the same dataset, thus from the same learner population and language.

We select four different prompts from POWERGRADING with a mean length between 3.3 and 4.8 tokens per answers and three different prompts from the ASAP dataset with an average length between 45 and 53 tokens and show the resulting learning curves for an instance-based setup in **Figure 7**. We observe that these prompts behave very differently despite a comparable length of the answers. Especially for the POWERGRADING data, performance with

[3] https://dkpro.github.io/dkpro-core/

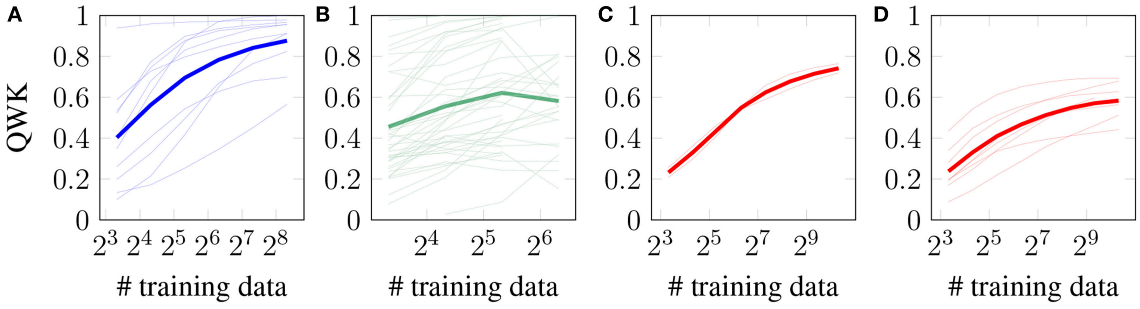


FIGURE 6 | Instance-based learning curves for datasets with different average lengths. Thin lines are individual prompts, while the thick line is the average for this dataset. **(A)** very short (POWERGRADING). **(B)** short (SRA). **(C)** medium (ASAP short). **(D)** long (ASAP long).

very few training data instances varies considerably showing other factors than length contribute to the performance. We assume that for these prompts (with often repetitive answers) the label distribution plays a role, as performance with few training instances suffers because chances are high that only members of the majority class are selected for scoring. For the ASAP prompts, those differences are less pronounced.

With the currently available data, we cannot make any claims about the influence of the prompt type itself, e.g., regarding domain (like *biology prompts are easier than literature prompts*) or modality of the prompt (as this would require having comparable prompts for example as listening and reading comprehension).

4.4. Language

In order to compare approaches solely based on the language involved, one would need the same prompts administered to comparable learner population but in different languages. The only such available datasets we know about are ASAP and ASAP-DE. ASAP-DE uses a subset of the prompts of ASAP translated to German and provides answers from German-speaking crowdworkers (Horbach et al., 2018). These answers were annotated according to the same annotation guidelines. So, while trying to be as comparable as possible, the datasets still differ in the learner population, in addition to the language. Horbach et al. (2018) compared instance-based automatic scoring on the two datasets and found results to be in a similar range with a slight performance benefit for the German data. However, they also reported differences in the nature of the data – resulting potentially from the different learner populations – , such as a different label distribution and considerably shorter answers for German, which they attribute to crowdworkers being potentially less motivated then school students in an assessment situation. Therefore, it is unclear whether any of those differences can be blamed on the language difference or the difference in learner population. More controlled data collections would be possible to get results that are specific to the language difference only. One such data collection with answers from students from different countries and thus various language backgrounds is the data from the PISA studies.[4] Such data would be an ideal testbed to compare learner populations with different native languages on the same prompt administered in various languages.

4.5. Learner Population

The results mentioned above for the different languages might equally be used as a potential example for the influence of different learner populations. In order to fully isolate the effect of learner population, one would need to collect the same dataset from two different learner groups such as native speakers vs. language learners or high-school vs. university students. To the best of our knowledge, such data is currently not available.

However, one aspect of different learner population is their tendency to make spelling errors. In experiments on the ASAP dataset, Horbach et al. (2017) found that the amount of spelling errors present in the data did not negatively influence content scoring performance. Only if the amount of spelling errors per answer was artificially increased, scoring performance decreased, especially, if errors followed a random pattern (unlikely to occur in real data) and if scoring methods relied on the occurrence of certain words and ignored sub-word information (i.e., certain character combinations).

4.6. Label Set

When discussion influence factors, we assumed that a dataset with more individual labels is harder to score than a dataset with binary labels. The influence of different label sets was already tested in previous work, especially in the SemEval Shared Task "The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge" (Dzikovska et al., 2013). The SRA dataset used for this challenge is annotated with three label sets of different granularity: two, three or five labels providing increasing levels of feedback to the learner. The two-way task just informs learners whether their answer was correct or not. The 3-way task additionally distinguishes between contradictory answers (contradicting the learner answers) and other incorrect answers. In the 5-way task, answers classified as incorrect in the 3-way task are classified in an even more fine-grained manner as "partially

[4] http://www.oecd.org/pisa/

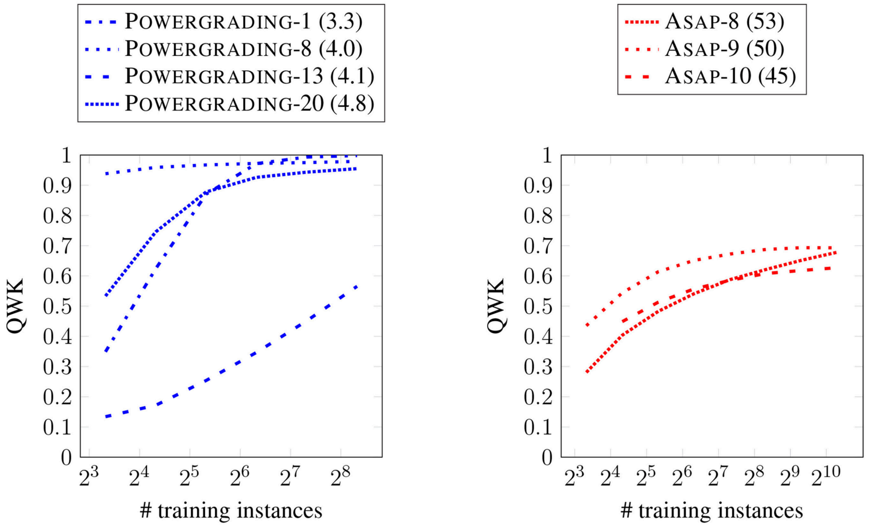


FIGURE 7 | Instance-based learning curves for POWERGRADING and ASAP prompts with comparable lengths.

correct, but incomplete," "irrelevant for the question," or "not in the domain" (such as *I don't know.*).

Seven out of nine systems participating the SemEval Shared Task reported results for each of these label sets. For all of them performance was best for the 2-way task (with a mean weighted F-Score of .720 for the best performing system) and worst for the 5-way task (0.547 mean weighted F-Score, again for the best performing system, which was a different one then for the 2-way result). This clearly shows that the expected effect of more fine-grained label sets being more difficult to score automatically.

5. CONCLUSION AND FUTURE WORK

In this paper, we discussed the different influence factors that determine how much variance we see in the learner answers toward a specific prompt and how this variance influences automatic scoring performance. These factors include the type of prompt, the language in the data, the average length of answers as well as the number of training instances that are available. Of course, these factors are interdependent and influence each other. It is thus hard to decide based on purely theoretical speculations whether, for example, medium length answers to a factoid question given by German native speakers annotated with binary scoring labels and with a large number of training instances are easier or harder to score than shorter answers in non-native English with numeric labels and a smaller set of training instances. Such questions can only be answered empirically, but the available datasets do not nearly cover the available parameter space exhaustively, so that such experiments are not possible in a straightforward manner. That makes it hard to compare different approaches in the literature and it is also a challenge to estimate the performance on new data. Therefore, we presented experiments that show the influence of some of the discussed factors on content scoring.

Our findings give researchers as well as educational practitioners hints about whether content scoring might work for a certain new dataset. At the same time, our paper also highlights the demand for more systematic research, both in terms of dataset creation and automatic scoring. For a number of influence factors, we were not able to clearly assess their influence because data that would allow to investigate a single influence parameter in isolation does not exist. It would thus be desirable for the automatic scoring community to systematically collect new datasets varying only in specific dimensions, such as to ask the same prompt to different learner populations and in different languages in order to further broaden our knowledge about the full contribution of these factors.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https: www.kaggle.com/c/asap-sas, https://www.microsoft.com/en-us/download/details.aspx?id=52397 and https://www.cs.york.ac.uk/semeval-2013/task7/index.php%3Fid=data.html.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work has been funded by the German Federal Ministry of Education and Research under grant no. FKZ 01PL16075.

REFERENCES

Anderson, L., Krathwohl, D., and Bloom, B. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman.

Andresen, M., and Zinsmeister, H. (2017). "The benefit of syntactic vs. linear n-grams for linguistic description," in *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, 4–14.

Bailey, S., and Meurers, D. (2008). "Diagnosing meaning errors in short answers to reading comprehension questions," in *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications* (Columbus), 107–115.

Bär, D., Biemann, C., Gurevych, I., and Zesch, T. (2012). "UKP: computing semantic textual similarity by combining multiple content similarity measures," in *Proceedings of the 6th International Workshop on Semantic Evaluation, Held in Conjunction With the 1st Joint Conference on Lexical and Computational Semantics* (Montreal, QC), 435–440.

Basu, S., Jacobs, C., and Vanderwende, L. (2013). Powergrading: a clustering approach to amplify human effort for short answer grading. *Trans. Assoc. Comput. Linguist.* 1, 391–402. doi: 10.1162/tacla00236

Beinborn, L., Zesch, T., and Gurevych, I. (2016). "Predicting the spelling difficulty of words for language learners," in *Proceedings of the Building Educational Applications Workshop at NAACL* (San Diego, CA: ACL), 73–83.

Bhagat, R., and Hovy, E. (2013). What is a paraphrase? *Comput. Linguist.* 39, 463–472. doi: 10.1162/COLI_a_00166

Burrows, S., Gurevych, I., and Stein, B. (2014). The eras and trends of automatic short answer grading. *Int. J. Art. Intell. Educ.* 25, 60–117. doi: 10.1007/s40593-014-0026-8

Burstein, J., Leacock, C., and Swartz, R. (2001). *Automated evaluation of essays and short answers*. Abingdon-on-Thames: Taylor & Francis Group.

Burstein, J., Tetreault, J., and Madnani, N. (2013). The e-rater automated essay scoring system. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, eds M. D. Shermis and J. Burstein (Routledge), 55–67.

Conort, X. (2012). "Short answer scoring: explanation of Gxav solution," in *ASAP Short Answer Scoring Competition System Description*.

Dagan, I., Roth, D., Sammons, M., and Zanzotto, F. M. (2013). *Recognizing Textual Entail-ment: Models and Applications*. Morgan & Claypool Publishers.

Daxenberger, J., Ferschke, O., Gurevych, I., and Zesch, T. (2014). "Dkpro tc: a java-based framework for supervised learning experiments on textual data," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (Baltimore, MD: Association for Computational Linguistics), 61–66.

Day, R. R., and Park, J. S. (2005). Developing reading comprehension questions. *Reading Foreign Lang.* 17, 60–73.

Dzikovska, M. O., Nielsen, R., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., et al. (2013). "Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge," **SEM 2013: The First Joint Conference on Lexical and Computational Semantics*.

Ellis, R., and Barkhuizen, G. P. (2005). *Analysing Learner Language*. Oxford University Press Oxford.

Fellbaum, C. (ed.). (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: Language, Speech, and Communication. MIT Press.

Firth, J. R. (1957). *A Synopsis of Linguistic Theory 1930-55.* London: Longmans.

Gale, W., Church, K. W., and Yarowsky, D. (1992). "Estimating upper and lower bounds on the performance of word-sense disambiguation programs," in *Proceedings of the 30th Annual Meeting on Association for Computational Linguistics* (Association for Computational Linguistics), 249–256.

Galhardi, L., Barbosa, C. R., de Souza, R. C. T., and Brancher, J. D. (2018). "Portuguese automatic short answer grading," in *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, 1373.

Haley, D. T., Thomas, P., De Roeck, A., and Petre, M. (2007). "Measuring improvement in latent semantic analysis-based marking systems: using a computer to mark questions about html," in *Proceedings of the Ninth Australasian Conference on Computing Education - Volume 66*, ACE '07 (Darlinghurst, NSW: Australian Computer Society, Inc.), 35–42

Heilman, M., and Madnani, N. (2015). "The impact of training data on automated short answer scoring performance," in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 81–85.

Higgins, D., Brew, C., Heilman, M., Ziai, R., Chen, L., Cahill, A., et al. (2014). Is getting the right answer just about choosing the right words? the role of syntactically-informed features in short answer scoring. *arXiv [preprint]. arXiv:1403.0801*.

Horbach, A., Ding, Y., and Zesch, T. (2017). "The influence of spelling error on content scoring performance," in *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications* (Taipei: AFNLP), 45–53.

Horbach, A., and Palmer, A. (2016). "Investigating active learning for short-answer scoring," in *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 301–311.

Horbach, A., Stennmanns, S., and Zesch, T. (2018). "Cross-lingual content scoring," in *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications* (New Orleans, LA: Association for Computational Linguistics), 410–419.

Jesensky, J. (2012). "Team JJJ technical methods paper," in *ASAP Short Answer Scoring Competition System Description*.

Landauer, T. K., and Dumais, S. T. (1997). A solution to plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104:211.

Meecham, M., and Rees-Miller, J. (2005). Language in social contexts. *Contemporary Linguistics* (Boston, MA: Bedford/St. Martin's), 537–590.

Meurers, D., Ziai, R., Ott, N., and Kopp, J. (2011a). Corpus of reading comprehension exercises in German. *CREG-1032, SFB 833: Bedeutungskonstitution - Dynamik und Adaptivität sprachlicher Strukturen, Project A4, Universität Tübingen.*

Meurers, D., Ziai, R., Ott, N., and Kopp, J. (2011b). "Evaluating answers to reading comprehension questions in context: results for german and the role of information structure," in *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, TIWTE '11 (Stroudsburg, PA: Association for Computational Linguistics), 1–9.

Meurers, D., Ziai, R., Ott, N., and Kopp, J. (2011c). "Evaluating answers to reading comprehension questions in context: results for german and the role of information structure," in *Proceedings of the TextInfer 2011 Workshop on Textual Entailment* (Edinburgh: Association for Computational Linguistics), 1–9.

Mohler, M., Bunescu, R. C., and Mihalcea, R. (2011). "Learning to grade short answer questions using semantic similarity measures and dependency graph alignments," in *ACL*, 752–762.

Mohler, M., and Mihalcea, R. (2009). "Text-to-text semantic similarity for automatic short answer grading," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics), 567–575.

Padó, U. (2016). "Get semantic with me! the usefulness of different feature types for short-answer grading," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2186–2195.

Padó, U. (2017). "Question difficulty–how to estimate without norming, how to use for automated grading," in *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 1–10.

Pado, U., and Kiefer, C. (2015). "Short answer grading: when sorting helps and when it doesn't," in *Proceedings of the 4th workshop on NLP for Computer Assisted Language Learning at NODALIDA 2015, Vilnius, 11th May, 2015* (Linköping University Electronic Press), 42–50.

Patil, P., and Agrawal, A. (2018). *Auto Grader for Short Answer Questions.* Stanford University, CS229.

Resnik, P., and Lin, J. (2010). "Evaluation of nlp systems," in *The Handbook of Computational Linguistics and Natural Language Processing.*

Reznicek, M., Ludeling, A., and Hirschmann, H. (2013). "Competing target hypotheses in the falko corpus," in *Automatic Treatment and Analysis of Learner Corpus Data*, Vol 59, eds A. Díaz-Negrillo, N. Ballier, and P. Thompson (John Benjamins Publishing Company), 101–123.

Ringbom, H., and Jarvis, S. (2009). *Chapter 7: The Importance of Cross-Linguistic Similarity in Foreign Language Learning*. John Wiley & Sons, Ltd.

Riordan, B., Horbach, A., Cahill, A., Zesch, T., and Lee, C. M. (2017). "Investigating neural architectures for short answer scoring," in *Proceedings of the Building Educational Applications Workshop at EMNLP* (Copenhagen), 159–168.

Scriven, M. (1967). "The methodology of evaluation," in *Perspectives of Curriculum Evaluation, AERA Monograph Series on Curriculum Evaluation*, Volume 1, eds R. Tyler, R. Gagné, and M. Scriven (Chicago, IL: Rand McNally), 39–83.

Tandalla, L. (2012). Scoring short answer essays. *ASAP Short Answer Scoring Competition System Description*. The Hewlett Foundation.

Zbontar, J. (2012). Short answer scoring by stacking. *ASAP Short Answer Scoring Competition System Description. Retrieved July*.

Zesch, T., Heilman, M., and Cahill, A. (2015a). "Reducing annotation efforts in supervised short answer scoring," in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 124–132.

Zesch, T., Heilman, M., and Cahill, A. (2015b). "Reducing annotation efforts in supervised short answer scoring," in *Proceedings of the Building Educational Applications Workshop at NAACL* (Denver, CO), 124–132.

Zesch, T., and Horbach, A. (2018). "ESCRITO - An NLP-Enhanced Educational Scoring Toolkit," in *Proceedings of the Language Resources and Evaluation Conference (LREC)* (Miyazaki: European Language Resources Association (ELRA)).

Ziai, R., Ott, N., and Meurers, D. (2012). "Short answer assessment: establishing links between research strands," in *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, NAACL HLT '12 (Stroudsburg, PA: Association for Computational Linguistics), 190–200.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 25 April 2019
doi: 10.3389/fpsyg.2019.00825

Multiple-Choice Item Distractor Development Using Topic Modeling Approaches

Jinnie Shin, Qi Guo and Mark J. Gierl*

Centre for Research in Applied Measurement and Evaluation, Department of Educational Psychology, University of Alberta, Edmonton, AB, Canada

Writing a high-quality, multiple-choice test item is a complex process. Creating plausible but incorrect options for each item poses significant challenges for the content specialist because this task is often undertaken without implementing a systematic method. In the current study, we describe and demonstrate a systematic method for creating plausible but incorrect options, also called distractors, based on students' misconceptions. These misconceptions are extracted from the labeled written responses. One thousand five hundred and fifteen written responses from an existing constructed-response item in Biology from Grade 10 students were used to demonstrate the method. Using a topic modeling procedure commonly used with machine learning and natural language processing called latent dirichlet allocation, 22 plausible misconceptions from students' written responses were identified and used to produce a list of plausible distractors based on students' responses. These distractors, in turn, were used as part of new multiple-choice items. Implications for item development are discussed.

Keywords: multiple-choice items, distractors, misconceptions, distractor generation, latent dirichlet allocation

OPEN ACCESS

Edited by:
Frank Goldhammer,
German Institute for International
Educational Research (LG), Germany

Reviewed by:
Qiwei He,
Educational Testing Service,
United States
Torsten Zesch,
University of Duisburg-Essen,
Germany

***Correspondence:**
Jinnie Shin
jinnie.shin@ualberta.ca;
eshin1@ualberta.ca

Specialty section:
This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Psychology

Received: *06 November 2018*
Accepted: *27 March 2019*
Published: *25 April 2019*

Citation:
Shin J, Guo Q and Gierl MJ (2019) Multiple-Choice Item Distractor Development Using Topic Modeling Approaches. Front. Psychol. 10:825. doi: 10.3389/fpsyg.2019.00825

INTRODUCTION

Multiple-choice testing is one of the most enduring and successful forms of educational assessment that remains in practice today. Multiple-choice items are used in educational testing because they permit the measurement of diverse types of knowledge, skills, and competencies (Haladyna, 2004; Downing, 2006; Popham, 2008). Multiple-choice items are efficient to administer; they are easy to score objectively; they can be used to sample a wide range of content; they require a relatively short time to administer (Haladyna, 2004; Haladyna and Rodriguez, 2013; Rodriguez, 2016). Downing (2006, p. 288), in his seminal chapter in the *Handbook of Test Development*, claimed that selected-response items, like multiple choice, are the most appropriate item format for measuring cognitive achievement or ability, especially higher-order cognitive skills, such as problem solving, synthesis, and evaluation. He also stated that this item format is both useful and appropriate for creating exams intended to measure a broad range of knowledge, ability, or cognitive skills across many domains.

Because of these important benefits, multiple-choice items continue to have broad appeal and, hence, application in education, despite some potential disadvantages, such as guessing effects and unintentionally exposing students' to wrong information. North American students take 100s of multiple-choice tests and answer 1000s of multiple-choice items as part of their educational experience. Chingos (2012) reported that one-third of the United States use multiple-choice items exclusively for assessing 4th grade and 8th grade students' math and reading skills. In higher education, a multiple-choice test is a common and widely used assessment format for measuring students' knowledge, especially in introductory courses with a large group of students. Multiple-choice testing is also used extensively for international assessments. In the 2015 administration of The Trends in International Mathematics and Science Study (TIMSS), for example, half of the mathematics and science items used the multiple-choice format (Mullis et al., 2016). In the 2015 administration of the Program for International Student Assessment (PISA), two-third of the items in reading, mathematics, and science assessments were multiple choice (OECD, 2016).

A multiple-choice item consists of a stem, options, and auxiliary information. The stem contains context, content, and/or the question the student is required to answer. The options include a set of alternative answers with one correct option and one or more incorrect options or distractors. Auxiliary information includes any additional content, in either the stem or option, required to create an item, including text, images, tables, graphs, diagrams, audio, and/or video. To answer a multiple-choice item, the student is presented with a stem and two or more options that differ in their relative correctness. Students are required to make a distinction among response options, several of which may be partially correct, in order to select the best or most correct option. Hence, the student must use her or his knowledge and problem-solving skills to identify the relationship between the content in the stem and the correct option. The incorrect options are called distractors because they are considered to be "distracting" to students with partial knowledge due to their plausibility to yield the correct option.

Creating multiple-choice items is a challenging task, particular when it comes to distractor development, because of the sheer volume of work that is required. For example, to create 100 multiple-choice items that consists of one correct option and four incorrect options, a content specialist has to create 100 stems and 100 correct options. The content specialist also needs to create 400 plausible but incorrect options. This challenge of distractor development is both daunting and, oftentimes, unsuccessful. Haladyna and Downing (1993) evaluated the distractors from four standardized multiple-choice tests. They evaluated the quality and plausibility of distractors based on the attractiveness of distractors. More specifically, they emphasized that plausible distractors should be able to attract more than 5% of the low-performing students, who failed to identify a correct answer. Based on such criteria, they found that only 8% of the items contained effective distractors.

To overcome the challenge of creating large numbers of effective distractors, researchers and practitioners have explored and implemented different strategies. The most common strategy focuses on a list of plausible but incorrect alternatives linked to common misconceptions or errors in thinking, reasoning, and problem solving (Haladyna and Downing, 1989; Case and Swanson, 2001; Vacc et al., 2001; Collins, 2006; Moreno et al., 2006, 2015; de la Torre, 2009; Tarrant et al., 2009; Rodriguez, 2011, 2016). Haladyna and Rodriguez (2013) in their textbook *Developing and Validity Test Items* claim that the most effective way to develop plausible distractors using misconceptions is to identify "common errors" elicited by a particular stem in the item prompt. These common errors serve as candidates for plausible distractors. Haladyna and Rodriguez state that common errors can be identified in two ways. First, they can be identified using the judgments of contents specialists who have a good understanding of teaching and learning within a specific content area and who can specify the common errors and misconceptions that arise when students learn a new topic or concept. Second, they can be identified by evaluating student answers to constructed-response item (i.e., an item that contains a stem by no options) where errors in reasoning, thinking, and problem solving are documented in the student's responses. The second approach—extracting student responses from constructed-response items—is the preferred strategy for identifying common errors because it is based on the actual response processes from students rather than the expected response processes inferred from the judgment of content specialists about how students respond to test items. However, identifying and extracting common errors and misconceptions from the actual response processes is a daunting task because large amounts of response data must be processes and this data, in turn, must be classified accurately in order to identify outcomes that could be used as distractors.

The purpose of this study is to introduce an augmented intelligence approach for systematically identifying and classifying misconceptions from the students' written responses that are pre-labeled for the purpose of creating distractors that can be used for multiple-choice items. Augmented intelligence is an area within artificial intelligence that deals with how computer systems can emulate and extend human cognitive abilities thereby helping to improve human task performance and to enhance human problem solving (Zheng et al., 2017). It requires the interaction between a human and a computer system in order for the system to produce an output or solution. Augmented intelligence combines the human capacity for judgment with the ability of modern computing using computational analysis and data storage to solve complex and, typically, unstructured problems. Augmented intelligence can therefore be used to characterize any process or system that improves the human capacity for solving complex problems by relying on a partnership between a human and a machine (Pan, 2016; Popenici and Kerr, 2017).

We introduce and demonstrate an augmented intelligence method that can be used for distractor development using latent dirichlet allocation (LDA; Blei et al., 2003). LDA is a statistical model used in machine learning and natural language processing which identifies specific topics and concepts within written texts. Specific words are expected to appear in a written text more or less frequently given a particular topic. LDA can be used to capture this expected outcome in a mathematical framework by focusing on the number of times words appeared in written text for different topics. Using LDA, content specialists can identify actual misconceptions based on students' response processes in order to create lists of plausible distractors.

Traditional Approach for Distractor Development

Distractors are one of the key components that affect the overall quality of multiple-choice items as well as the item's statistical characteristics (Gierl et al., 2017). Distractors are intended to distinguish between students who have not yet acquired the knowledge necessary to answer the item correctly from those who understand the content. Therefore, distractors in a multiple-choice item are designed to contain plausible but incorrect answers based on students' common errors or misconceptions so that the option can measure students' level of mastery in a specific content area (e.g., Case and Swanson, 2001; Ascalon et al., 2007; Hoshino, 2013; Towns, 2014; Lai et al., 2016). Creating distractors using common errors and misconceptions result in multiple-choice items with increased diagnostic value as well as higher item quality (Haladyna and Downing, 1989; Case and Swanson, 2001; Briggs et al., 2006; Moreno et al., 2006, 2015; de la Torre, 2009; Tarrant et al., 2009; Rodriguez, 2011, 2016).

Haladyna and Rodriguez (2013) claimed that common errors and misconceptions could be identified using two different approaches. In the first approach, content specialists create individual distractors by hand that contain these common errors and misconceptions. Collins (2006) recommended that content specialists mimic students' problem solving processes by answering questions such as, "what is a common error for solving this problem?" and "what do students usually confuse this concept or idea with?" in order to identify plausible distractors. The most appealing aspect of this method lies in its practicality and ease of implementation. The distractors are created by content specialists familiar with the students and the content area to mimic the typical and the commons problems that are most likely to occur. While this approach is feasible, it is also based on three assumptions. First, plausible algorithms, rules, or sources of information can be specified by content specialists. Second, plausible but incorrect distractors can be produced using these sources. Third, the misconceptions identified by the content specialists from these sources are, in fact, the same misconceptions held by the students. Proper alignment of the assumptions is critical for creating distractors that measure students' actual errors and misconceptions. Moreover, the alignment must occur for each distractor across every multiple-choice item. Using our earlier example, if a content specialist writes 100 multiple-choice items and each item contains five options (i.e., one correct option and four distractors), then the content specialist must identify 400 plausible but incorrect alternatives that satisfy these three assumptions.

In the second approach, students' responses from existing constructed-response items are evaluated to identify common errors and misconceptions. That is, content specialists review students' responses from constructed-response items to identify mistakes, errors, and misunderstanding and then classify these outcomes to create a compiled list of plausible distractors (e.g., Bekkink et al., 2016). This approach addressed the inferential problem associated with the previous approach because it is based on actual student response data rather than judgments about expected response processes. In other words, approach two is data driven. Common errors and misconceptions identified using approach two come from the algorithms, rules, or sources of information used by students to produce incorrect answers. Unfortunately, the second approach suffers from the problem of practicality and ease of implementation because it is neither practical nor easy to use. As it is currently implemented, approach two is daunting because it entails a comprehensive review of students' written responses using a manual process with the goal of identify common errors and misconceptions that occur consistently and systematically. It is also a process fraught with interpretive problems because identifying common errors and misconceptions that occur systematically can be a subjective task (e.g., what are the characteristics of a systematic misconception). And, despite the potential benefits of using a data-driven approach, practically also dictates that the item development process should be relatively quick and efficient, even when large number of multiple-choice items are required. This requirement is challenging to address using the second approach, especially when large amounts of written text are available from a constructed-response item.

To-date, limited research has been conducted to investigate the application of augmented intelligence for the purpose of distractor development. Researchers have explored the significance of using students' misconceptions and common errors to create distractors. The approach used in these studies was based on identifying misconceptions using students written or verbal responses that, in turn, were manually categorize by content specialists to identify common errors and misconceptions (e.g., Vacc et al., 2001; Haladyna and Rodriguez, 2013; Moreno et al., 2015; Bekkink et al., 2016; Rodriguez, 2016). As noted earlier, a data-drive approach using students' responses is inherently beneficial for identifying the actual errors and misconceptions that students use when they produce incorrect answers. But it is also inherently limited because it is excessively time consuming and labor intensive to identify and classify errors from written text using a manual review process. To overcome this limitation, we introduce and illustrate a data-driven method for creating distractors based on student's common errors and misconceptions using LDA.

Topic Modeling and Latent Dirichlet Allocation

Locating keywords and topics to understand text is a simple and effective way for humans to classify textual information. To gather information about certain topics, for example, we often start from generating one or two key words to locate relevant documents that share common topics. Unfortunately, this approach quickly becomes unmanageable for humans when the amount of textual information begins to increase. For example, having content specialists manually review 1000s of students' responses to identify and then categorize common errors would be a time consuming and inefficient classification exercise.

To overcome this clustering challenge, topic modeling has been developed and used with machine learning and natural language processing algorithms to uncover the hidden topics in a document (Blei, 2012). These hidden topics can be identified without any pre-labeling, which means that topic models do not require pre-categorized or topic-labeled documents. In machine learning, these problems are described as an unsupervised learning approach, which means the structure of the problem includes targets or outputs which are unknown and hence the primary focus of learning is to understand the structure of the data. Therefore, in topic modeling, we attempt to identify hidden or unobserved target, topics, using the fully observed information, words.

If we assume that a sequence of words in a document is governed by the same unobserved topic, then we could simply compute the likelihood of a document to represent certain topic to determine the underlying topic of a document in an unsupervised setting. To find the common topics, topic modeling uses word occurrence information where certain words are expected to appear in a document more or less frequently depending on a particular topic. LDA is a generative probabilistic topic modeling algorithm (Blei et al., 2003), where each document is perceived as a mixture of several topics. Generative models take the information of how observed data was generated into account to build a model. Suppose, for instance, we have documents that were generated by complex procedures that are unknown.

Latent dirichlet allocation attempts to synthesize an approximated generation procedure and observed information (i.e., words) to uncover hidden topics, without any labels. Moreover, unlike other topic modeling approaches, LDA can not only produce interpretable topics and can handle unseen documents to assign topics. The generative process of LDA consists of three layers of sampling a topic distribution, sampling topics, and sampling words over topics. For example, after the number of words (or document length) and the number of topics are decided, a topic distribution is specified (e.g., 40% biology, 30% kinetics, and 30% psychology). Next, a topic is picked based on the topic mixture distribution and a word is picked based on the distribution over words corresponding to the topic. This process is then repeated until all the words are generated for each documents. **Figure 1** describes a graphical representation of the generative process of LDA.

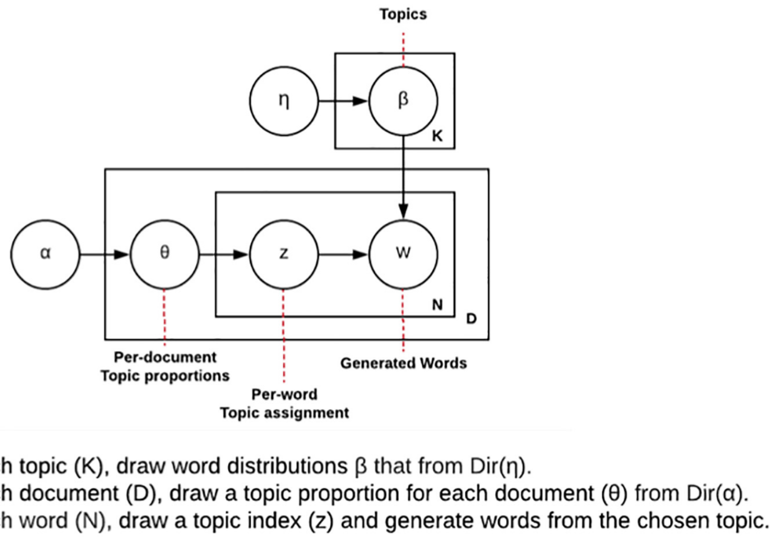
Given this process, LDA attempts to explore the hidden topics in a document by computing a posterior distribution of the hidden variables given a document. Due to a large number of possible topic structures, computing the probability of certain words under a specific topic (i.e., the distribution over words corresponding to the topic) becomes impossible to compute. To address this problem, LDA uses a method called Gibbs sampling (Porteous et al., 2008) where each word is randomly assigned in the document to one of the topics, which will provide the initial guess of the word-topic and word-document distribution. LDA assumes that all topic assignments except for the current word in question are correct, and then updates the assignment of the current word. This process is repeated to improve the assignment until a steady state is reached. Once the final assignment is identified, it is used to estimate the topic mixtures of each document.

Model Evaluation and Augmented Intelligence

While topic models can be used to extract meaningful and interpretable topic assignments, evaluating the final assignment is challenging using an unsupervised approach (Chang et al., 2009). Unsupervised learning tasks do not include pre-labeled targets. Instead human judgment is required to evaluate the practicality and usefulness of the topic modeling performance (Konrad, 2017). For example, the practicality of the topic model could be evaluated using the "human-in-the-loop" augmented intelligence approach, where humans are asked to locate a randomly substituted word or topic (Chang et al., 2009). If the human can reliably tell which one is a random intruder, then we can say that the trained topic yields a coherent and discernible topic (Chang et al., 2009). In addition, intrinsic measures (i.e., statistical measures) should also be considered for model evaluation. Such measures help evaluate how well the model fits the observed data.

Log-likelihood evaluates the probability of the observed data, given the model (Griffiths and Steyvers, 2004). Thus, we can locate the best model by attempting to produce the highest log-likelihood measure. The Kullback-Leibler (KL) divergence measure focuses on measuring the divergence among the topic distributions. KL divergence explicitly focuses on evaluating how much information we lose when we choose a certain model, by computing the symmetric KL divergence between the distribution of variance in the topic-word distribution and the marginal topic distribution (Cao et al., 2008; Arun et al., 2010). Thus, the best model can be determined by locating the point where the KL divergence measure reaches the lowest value (Arun et al., 2010).

Previous research has been conducted to demonstrate the usefulness of LDA for different types of topic modeling assignments. In education, for example, LDA has been used to uncover topics for essay scoring purposes (Meisner, 2018), implementing course recommendation systems (Apaza et al., 2014), and evaluating teachers (Moretti et al., 2015). However, to our knowledge, LDA has never been used to identify students' errors and misconceptions for the purpose of creating distractors that could be used to create multiple-choice items. Therefore, the purpose of the study is to describe a method for creating distractor by identifying students' misconceptions using the LDA topic modeling approach. Unlike the traditional approach where content specialists were responsible for using their judgments to analyze and evaluate students' responses in order to identify plausible misconceptions for distractors development, the current study provided a systematic and data-driven method to cluster students' written responses with similar underlying concepts in order to locate common mistakes. Once clustered, these responses become the basis for creating plausible distractors.



Step 1. For each topic (K), draw word distributions β that from Dir(η).
Step 2. For each document (D), draw a topic proportion for each document (θ) from Dir(α).
Step 3. For each word (N), draw a topic index (z) and generate words from the chosen topic.

FIGURE 1 | A conceptual representation of latent dirichlet allocation (LDA).

MATERIALS AND METHODS

Data

An open source data set collected and released from the short-answer scoring competition called Automated Student Assessment Prize (ASAP) was used in the study[1]. As the data set is publicly available, ethical approval was not sought in the study. ASAP was held in 2012. The competition was designed to promote the capabilities of effective scoring system using automated essay scoring frameworks and to provide efficient classroom essay scoring tools for practitioners. The competition included two phases. The first phase focused on developing robust automated scoring frameworks for relatively long responses (up to 650 words). The second phase focused on scoring short responses (up to 50 words). Both the competitions significantly contributed to promoting open and rigors model development for automated essay scoring (Shermis, 2014, 2015).

For the short-essay scoring competition, 10 data sets were released and each data set was generated from a single prompt. The responses were produced by students in grade 10. Each data set was based on a unique prompt in different disciplines, such as Language Arts, Biology, and Science. All the responses were pre-labeled, scored by two human-raters. The current study used data set six from Biology to demonstrate the proposed method. This data was chosen to demonstrate the proposed method for three reasons. Fist, the current method requires pre-labeled data set and the data set six consisted of the resolved-score (or final score) based on the agreement of the two human raters. Second, the prompt required students to respond using multiple answers thereby producing a variety of diverse responses from a single prompt. In addition, the original constructed-response prompt could be easily reformatted into a multiple-choice stem.

More specifically, we used 1,515 responses from the original training set, where students were asked to list and describe three processes used by cells to control the movement of substances across the cell membrane (see **Appendix A**). The particular number of training responses were selected based on the score assigned by two independent human raters. The final score corresponded to the number of correctly identified answer and we only selected the responses where students failed to identify any correct answer (i.e., score 0), as the focus of this study is on extracting common errors and misconceptions.

Distractor Development Stage 1: Data Preparation

To achieve clear and interpretable clusters of topics, pre-processing is required. First, all of the misspelled words were corrected. Second, words were converted into lower cases and lemmatized using the Python NLTK library (Bird et al., 2009). Lemmatization is the process of grouping the words together so they can be analyzed as a single item based on their dictionary form. For example, the words 'studies' and 'studying' would be lemmatized into 'study.' Third, digits, non-alphabetic words (e.g., #, %, &, @), and stop words (e.g., a, and, but, how) were removed and all punctuation was specified as a separate word. Fourth, responses were separated into sentences allowing each sentence to be denoted as a separate topic.

Pre-processing is also focused on spelling correction using a combination of several approaches. We used the word embedding-based model for spelling correction. Word embedding-based models use the semantic similarities of words to determine the best candidate of a misspelled word (Nagata et al., 2017, see **Appendix C**). We used a list of words provided in the pre-trained GloVe embedding (Pennington et al., 2014), which were trained on six billion words from Wikipedia 2014 and Gigaword 5. We attempted to locate the best candidate of an incorrect word from the Glove embedding word list based on a cosine-similarity score. Using the embedding-based spell

[1] www.kaggle.com/c/asap-sas

correction, we could successfully correct more than 95% of the misspelled words, while some of the remaining misspelled words that could not be fixed with the methods were correctly manually. This approach was chosen after attempting existing spell checkers in Python and the correction results were relatively lower than expected (e.g., NLTK edit-distance with 78% correction). Such cases often included words that were significantly malformed, thus, providing very limited resemblance with a correct form.

Distractor Development Stage 2: Topic Clustering and Cluster Evaluation

The LDA model was constructed using the Python library lda 1.0.5. To generate clear and interpretable clusters of topics, model training and evaluation took place simultaneously. To enable flexible and robust learning, it is necessary to identify the ranges of several model parameters so the model with the optimum range can be identified. For example, the number of topic groups must be specified before training begins. The number of Gibbs sampling iteration must also be specified to train the model. To begin, the number of topics and sample iterations ranged from 1 to 50 and up to 800 iterations, respectively. These ranges were selected so that we can extract as many potential misconceptions as possible with a stable estimation. We set our initial range of the number of topics as a relatively large number, 50, so that the model could conduct a comprehensive categorization of common errors and misconceptions. In terms of the number of iterations, we evaluated the negative log-likelihood of the model at every 10 iterations and inspected whether a significant decrease or increase in log-likelihood occurred. The significance was evaluated based on a chosen tolerance value of 0.5. The results indicated that log-likelihood stabilized around 800 iterations. The performance of our initial model was evaluated using the perplexity measure. Perplexity is a commonly used topic-model measure that is computed by dividing a negative log-likelihood by the number of words (see **Appendix C**). As the name suggests, perplexity provides the degree of 'uncertainty' or 'confusion' the model has in assigning probabilities to text. Therefore, we could determine the optimal number of topics by locating the model with the lowest perplexity.

Then, the topic clusters were visualized to evaluate the clustering. Topic clusters were projected in a two-dimensional space by computing the distance between topics using t-distributed stochastic neighbor embedding (t-SNE). T-SNE is a dimensionality reduction algorithm for high-dimensional data visualization. The idea of t-SNE is to find a probability distribution that is a function of the smallest number of coordinates and to create a similar distribution function to reduce the dimensionality. Assume that we want to calculate the probability of finding two points i and j at the squared Euclidean distance between the points, $||x_i - x_j||^2$. T-SNE attempts to match the distribution using a Student's-t distribution, while attempting to learn the y coordinates of the points (i.e., y_i and y_j) in the lower dimension. If the visualized clusters are significantly overlapping and malformed, then the number of topics should be adjusted. In addition, the KL divergence was used as an evaluation criterion for the visualization because it helps determine the similarity of the two distributions. The learning algorithm attempts to create a clear visualization of distinctive topic clusters while minimizing KL divergence to locate the optimal model. To do so, several adjustments were necessary to determine the number of iterations, the learning rate, and the perplexity rate. While the number of iterations and the learning rate determines the efficiency and accuracy of model learning through controlling for the weight adjustments, the perplexity rate controls for the effective number of cluster neighbors. Finally, interpretability of the clusters was evaluated by summarizing the clustered sentences using the Python library genism summarization. Gensim summarization conducts a text rank-based summarization using a variation of the TextRank algorithm (Barrios et al., 2016). TextRank attempts to construct a graph from a document, where sentences (or nodes) are connected with each other via edges. Edges represent the similarity between the sentences, which are often computed based on the word overlap between the two sentences. TextRank hypothesizes that the most important sentence in a text as the one that is the most frequently connected in a graph. We chose this approach as previous studies have demonstrated relatively good performance using the method, while it does not require any manual annotation (Mihalcea and Tarau, 2004). The summaries were created so that content specialists could effectively evaluate the plausibility of the extracted common errors and misconceptions.

In the study, we refer to content specialists as the experts who are experienced in item writing in particular subjects. With this type of content expertise, validating the plausibility of summarized common errors and misconceptions could improve the quality of distractors which are generated from each topic cluster. To do so, content specialists could discuss and attempt to identify where each misconception originated from. For example, if the content of a cluster includes morphologically or phonetically similar words with correct answers, the specialists could conclude that the misconception originated from the confusion in recalling certain terminologies or associating a term with a correct definition. Also, content specialists could be encouraged to answer more concrete questions to evaluate the quality of clusters. Such questions could include, "How many of the clusters do you find meaningful?" and "Is the cluster describing a commonly well-identified misconception regarding the topic?" This would help content specialists to evaluate distractors thoroughly, while providing important information to evaluate the capacity of the current system.

Distractor Development Stage 3: Item and Distractor Formation

In stage 3, content specialists formulate distractors using the common errors and misconception clusters identified in the previous stage. We propose several methods that could promote more systematic distractor development using students' misconceptions. The distractor generation process can be distinguished based on the question type (or stem) that content specialists pose regarding a topic. First, the content specialists could decide to change the format of the original question from the constructed-response item to a multiple choice item format, while attempting to measure the same construct of interest (e.g., which of the following procedures is correct about cell movement?). In this case, we could use the cluster summarizations and the key words and phrases directly. In stage 2, we explored how each misconception cluster can be represented using key words and summarization. Thus, using key words or summarized sentences as distractors would be able to attract students with different levels of understanding effectively. Alternatively, content specialists could develop a question that focuses on specific sub-concepts of a topic. Active- or passive-transport could be good examples of sub-concepts to evaluate, that is closely associated with the original question. In this case, distractors could be directly located based on students' responses from the cluster, where students appeared to have trouble understanding the concepts of active- and passive-transport. We will present how the two methods can be utilized more thoroughly using examples in the next section.

Generating distractors using students' misconceptions have been identified as one of the most effective way in developing multiple-choice items (Haladyna and Rodriguez, 2013). However, with our augmented intelligence approach, which require content specialists' judgment in the evolution process, we believe the effectiveness of distractors could still significantly depend on the content specialists judgments. Therefore, while we encourage further studies on the effectiveness of the distractors generated using the proposed methods, it was out of our scope of research to provide empirical results on behaviors of distractors in a real test setting. We will discuss such concerns more thoroughly in the limitation section with several suggestions for future research.

RESULTS

Topic Clustering and Cluster Evaluation Results

In the original constructed-response item, students were asked to provide three correct responses to the following item: "List and describe three processes used by cells to control the movement of substances across the cell membrane." The results indicated that the optimal LDA model identified 22 common misconceptions. The number of topic clusters were selected based on the log-likelihood measure as well as the KL divergence. The model achieved a perplexity of 34.76 after 800 iterations and the lowest KL divergence of 40.50 with 22 topics. As discussed earlier, the log-likelihood measure provides the probability of the observed data given the model (Griffiths and Steyvers, 2004).

In addition, the interpretability and plausibility of each topic cluster was evaluated using extracted key words and summaries. A full list of topic key words and summaries can be found in **Appendix B**. Six to eight topic key words were used for each topic cluster. They were chosen based on the strength of association to represent the topic cluster and the strength was measured by weights assigned to each word. In addition, summaries were generated for each cluster to increase their interpretability. This information was designed to help the content specialists to interpret students' common errors and misconceptions and to evaluate the representativeness of the clusters to form plausible distractors. For example, topic 20 included several key words, such as 'mRNA,' 'RNA,' 'tRNA,' 'DNA,' 'information,' 'translation,' 'transcription,' and 'messages.' Content specialists formed their initial impression on each misconception based on these key words. In addition, by reading the summary which states "mRNA carries messages from the nucleus to other organs tRNA transports DNA to places with in the cell rRNA," content specialists can understand specific contexts and associations among the key words more thoroughly so they can make more informed decision about whether the cluster could be used to create a plausible distractor which represents a common error or misconception.

Item and Distractor Formation Results

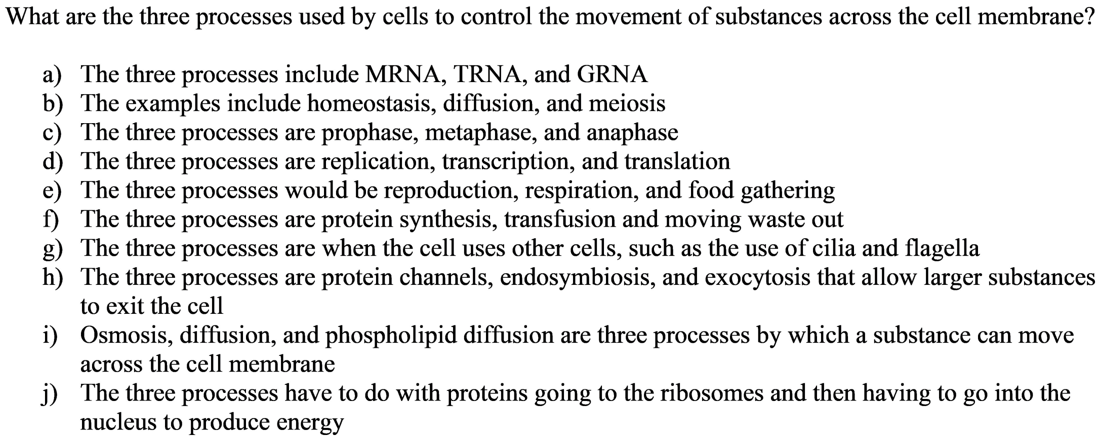
A set of distractors were generated using the evaluated clusters of students' common errors and misconceptions. In addition to create distractors for the originally proposed item, where students were required to describe three processes used by cells to control the movement of substances across the cell membrane, we explored the capacity of the current method in generating distractors on additional cluster-specific items. The following examples introduce a step-by-step breakdown of the distractor generation procedures.

Example 1: Generating Distractors for the Original Prompt

As shown in **Figure 2**, a multiple-choice item was created from the original constructed-response item. Reflecting the original prompt, the stem was changed to "What are the three processes used by cells to control the movement of substance across the cell membrane?" To generate distractors that could each reflect different common error and misconception, the list of options was created by locating students' responses with key words from the stem, such as 'processes,' 'movement,' or 'substances' from each misconception topic cluster. More specifically, the option *g* represents the cluster 13 (see **Appendix B**), where students describe the movement of flagellum as part of the movement of substances across the cell membrane. In this example, the correct answer is i, while the other options were produced to represent students' misconceptions.

Example 2: Generating Distractors Using Additional Prompts

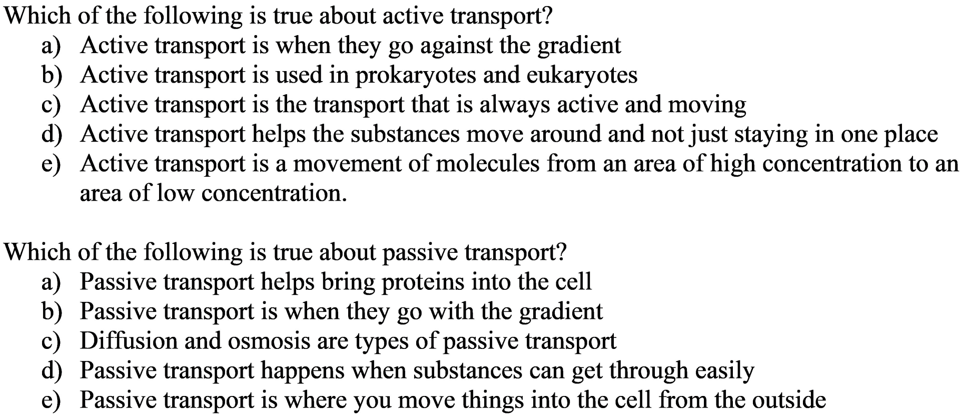
As shown in **Figure 3**, the proposed method could be extended to generate distractors for cluster-specific items. Cluster-specific items refer to items that are generated to further evaluate students' understanding that reflect the misconceptions captured in a particular content cluster. For example, **Figure 3** introduces two cluster-specific items, which were posed based on students' responses in cluster 2 (see **Appendix B**). In cluster 2, students had trouble correctly explaining and distinguishing between the two concepts of active and passive transports. Therefore, to evaluate students' understanding on active and passive transport, two additional multiple-choice stems were created: "Which of the following is true about active transport?" and "Which of the following is true about the passive



What are the three processes used by cells to control the movement of substances across the cell membrane?

a) The three processes include MRNA, TRNA, and GRNA
b) The examples include homeostasis, diffusion, and meiosis
c) The three processes are prophase, metaphase, and anaphase
d) The three processes are replication, transcription, and translation
e) The three processes would be reproduction, respiration, and food gathering
f) The three processes are protein synthesis, transfusion and moving waste out
g) The three processes are when the cell uses other cells, such as the use of cilia and flagella
h) The three processes are protein channels, endosymbiosis, and exocytosis that allow larger substances to exit the cell
i) Osmosis, diffusion, and phospholipid diffusion are three processes by which a substance can move across the cell membrane
j) The three processes have to do with proteins going to the ribosomes and then having to go into the nucleus to produce energy

FIGURE 2 | An example question and distractors generated for the original prompt.



Which of the following is true about active transport?
a) Active transport is when they go against the gradient
b) Active transport is used in prokaryotes and eukaryotes
c) Active transport is the transport that is always active and moving
d) Active transport helps the substances move around and not just staying in one place
e) Active transport is a movement of molecules from an area of high concentration to an area of low concentration.

Which of the following is true about passive transport?
a) Passive transport helps bring proteins into the cell
b) Passive transport is when they go with the gradient
c) Diffusion and osmosis are types of passive transport
d) Passive transport happens when substances can get through easily
e) Passive transport is where you move things into the cell from the outside

FIGURE 3 | Example questions and distractors generated for the sub-topics of the original prompt.

transport?" To generate distractors for the cluster-specific items, we implemented the same process where the key words and phrases (i.e., active transport, passive transport) were used to locate students' responses that included these key terms. Unlike the first example, the distractors were only located among the responses in cluster 2 as the items were created based on cluster 2. The correct option is *a* and *b*, respectively.

DISCUSSION

The recent introduction of different applications of augmented intelligence in educational assessment have brought about dramatic changes in the field by promoting efficient new test development and administration procedures (Popenici and Kerr, 2017). Augmented intelligence, which is a branch of artificial intelligence, helps content experts broaden their capabilities and make more informed decision in a timely manner with appropriate technological support. For instance, with a machine-aided scoring system, experts can score essays more efficiently because the machine can be used to help distinguish problematic essays that fail to map onto a scoring rubric from more coherent essays. Currently, little research has been conducted to investigate the application of augmented intelligence in item development, especially as it relates to creating distractors. Effective distractors can attract students with a partial understanding, in other words, discriminating students who have not yet reached the mastery level of comprehension regarding the concept. Thus, generating effective distractors is directly associated with increasing the quality of an item and its characteristics (i.e., item difficulty and discrimination; DiBattista and Kurzawa, 2011). Studies have been conducted to explore the significance of using students' misconceptions and common errors to create distractors (e.g., Vacc et al., 2001; Moreno et al., 2015; Rodriguez, 2016). Misconceptions are typically gathered using students written or verbal responses on similar or connected topics and content experts manually categorize and identify plausible misconceptions using the written response evidence (Bekkink et al., 2016). In other cases, content experts attempt to mimic students' thought processes in order to identify plausible errors

(Haladyna and Rodriguez, 2013). However, these approaches are unfeasible when large numbers of items must be created. To overcome this limitation, we introduced and illustrated a data-driven method for generating distractors based on misconceptions from students' written responses using the workflow presented in **Figure 4**.

It is important to acknowledge that the current methods attempt to incorporate both machine- or data-driven and experts-driven approaches harmoniously in every stage. While the data-driven approach provides prominent benefits in facilitating a systematic and effective distractor generation process, we believe the intervention from experts could help improving the system, behaving as a gatekeeper for quality insurance of the final product, distractors. Especially in educational assessments, content experts' decisions are often considered a reference or gold-standard in making the ultimate high-stakes decisions. The steps in **Figure 4** workflow were used to identify 22 distinct clusters of common errors and misconceptions using students' written responses from a constructed-response item in Biology. In the first data processing stage, we primarily used the data-driven approach to pre-process the responses (e.g., lemmatization, tokenization, remove punctuations, and non-alphabetic words). Also, while we corrected the majority of misspelled words using the embedding-based approach, it was still required to conduct a few manual corrections. In the response analysis stage, clusters were created automatically using a topic-modeling approach, then, content experts were required to evaluate the interpretability and plausibility of the extracted clusters, the information was used to generate a list of 22 plausible distractors that, in turn, helped create a parallel multiple-choice item. A parallel multiple-choice item refers to an item originally presented as a constructed-response task that has been reformatted into a selective-response task. The quality of generated distractors can be further empirically evaluated by pilot testing in a classroom evaluation setting and we will discuss more details about the evaluation of item characteristics in the next section.

Implications for Future Research

The current study has implications for distractor writing practices, specifically, and item development, more generally. Topic modeling allows content experts to use student responses in a more adaptive and productive way. Written responses represent an enormous source of valuable information about students' understanding, which is not only related to the construct of interest, but also to misconceptions about that construct. To-date, little effort has been spent exploring the use of machine learning methods for gathering and using information about misconceptions that can be found is students constructed responses. Using the method described and illustrated in this study, researchers and practitioners

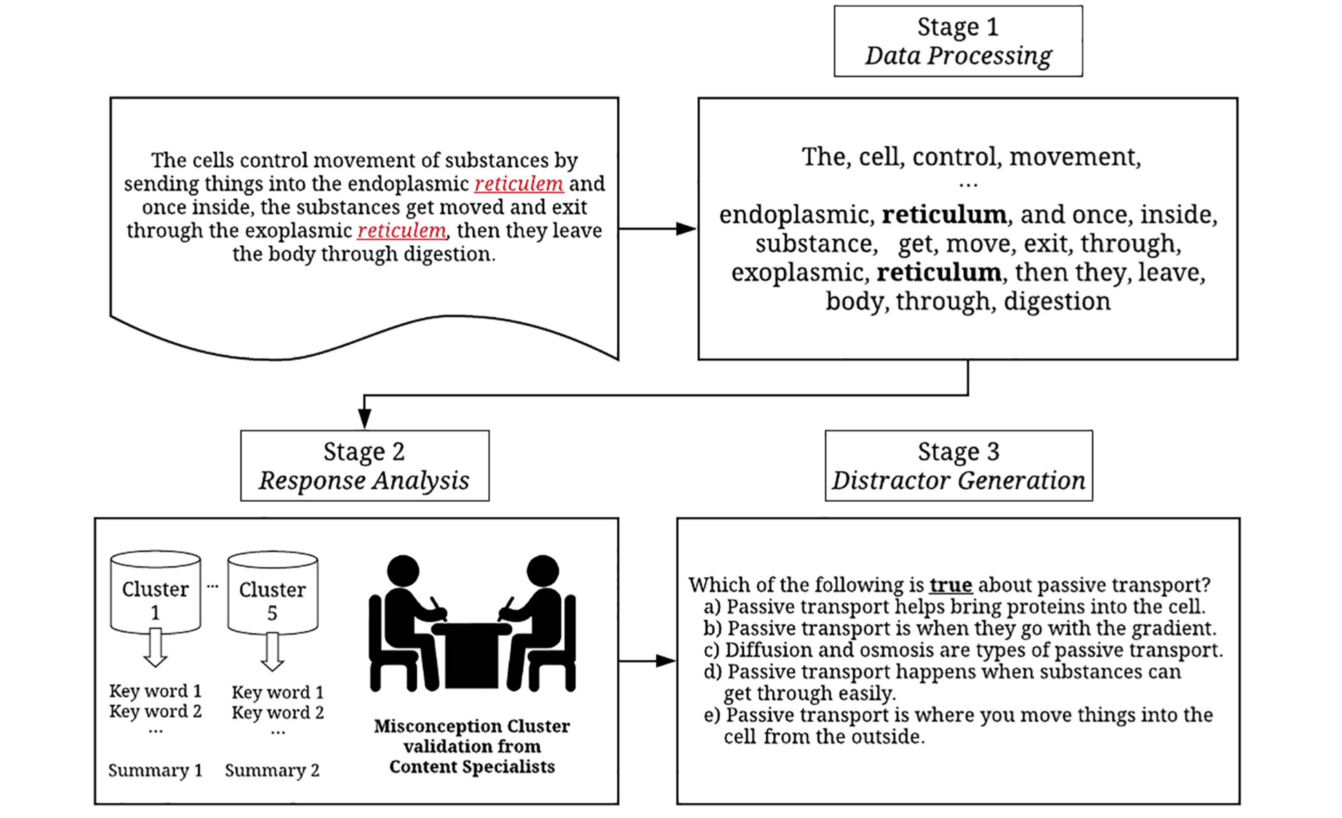

Stage 1
Data Processing
The cells control movement of substances by sending things into the endoplasmic reticulem and once inside, the substances get moved and exit through the exoplasmic reticulem, then they leave the body through digestion.
The, cell, control, movement,
...
endoplasmic, reticulum, and once, inside, substance, get, move, exit, through, exoplasmic, reticulum, then they, leave, body, through, digestion
Stage 2
Response Analysis
Cluster 1
...
Cluster 5
Key word 1
Key word 2
...
Summary 1
Key word 1
Key word 2
...
Summary 2
Misconception Cluster validation from Content Specialists
Stage 3
Distractor Generation
Which of the following is true about passive transport?
a) Passive transport helps bring proteins into the cell.
b) Passive transport is when they go with the gradient.
c) Diffusion and osmosis are types of passive transport.
d) Passive transport happens when substances can get through easily.
e) Passive transport is where you move things into the cell from the outside.

FIGURE 4 | A comprehensive framework of the distractor generation process.

can now use the written responses gathered in assignments and tests to plan future lessons and to create more student-adapted learning activities and assessments. The method can also be used to provide evidence for students' developmental level of understanding about certain concepts. For example, by analyzing the responses from the higher-ability group and compare the misconception clusters with the ones from the lower-ability group, more in-depth information can be gathered to create a comprehensive picture of how students' level of understanding develops on specific concepts and within specific content areas.

Distractor Development and Item Generation

Potentially the most important future application of this method resides in its application to automatic item generation (AIG; Irvine and Kyllonen, 2002; Gierl and Haladyna, 2013). AIG is a relatively new but rapidly evolving research area where cognitive and psychometric modeling practices guide the production of tests that include items generated with the aid of computer technology. Gierl and Lai (2013, 2016) developed a three-step process for AIG. In step 1, content specialists create a cognitive model for AIG.

Currently, distractor development poses a unique and consequential problem in AIG in the step 2 item modeling stage. For the selected-response format, items must not only include a stem with a corresponding correct option, but also include a set of distractors. Distractors in AIG are typically designed from a list of plausible but incorrect alternatives linked to misconceptions identified by content specialists. Because AIG produces 100s of items, strategies are needed to create a correspondingly large number of plausible but erroneous distractors. Distractor development for AIG is now guided by the distractor pool method with random selection (Gierl and Lai, 2016). To identify the content for the distractors, content specialists identify a list of plausible but incorrect options that are appropriate for all possible items generated with a given item model. Then, distractors are randomly selected from this pool of plausible but erroneous content and added to each generated item. This method is based on the assumption that a pool of plausible distractors can be created. A sample of these plausible distractors are selected at random to complete the item generation process. The strength of this method is its simplicity. This method can yield large numbers of distractors. The weakness of this method resides with the strong assumption that all pooled distractors are equally plausible and appropriate for all generated items. Equal plausibility and appropriateness is strong and, in many cases, restrictive assumption. Also, there is little reasoning to guide how distractors are paired with the correct option because pairing is achieved with random assignment.

To improve the plausibility and appropriateness of the distractors, rules, and rationales that yield errors or misconceptions can be used to create distractors. Distractor rationales are short descriptions that specify the reasoning which underlies each option. These rationales are currently provided by content specialists. But the rules can also be created using the method presented in our study to produce distractors that conform to specific, empirically-based, student misconceptions. Hence, distractors can be created systematically so that each distractor matches a rationale. This proposed approach could be called the *systematic generation with rationales method*. It would be based on the assumption that algorithms, rules, and procedures can first be articulated by content specialists and then used to create plausible but incorrect alternatives linked to students' actual misconceptions or errors in thinking, reasoning, and problem-solving. The strength of this method is that the distractors are much more specific and, hence, plausible and appropriate, especially when compared to the distractor pool method with random assignment. Hence, integrating the outcomes from the topic modeling methods presented in this paper with new developments in AIG should be considered an important area of future research.

Limitation and Future Research

Even though the study was carefully designed and structured to minimize potential error with results and further interpretations, we found the three key limitations that should be addressed and carefully considered for future research: the main purposes of our study were to introduce a novel method of identifying students' misconceptions in a systematic manner to encourage efficient distractor generation for multiple-choice item development. Thus, our study could not investigate the item behaviors with generated distractors in a real test setting. Investigating the item behaviors in relation to the distractor quality would help us further understand the importance of item development with well-performing distractors. For example, DiBattista and Kurzawa (2011) demonstrated how the plausibility of distractors significantly affects item characteristics (e.g., item discrimination) in classroom assessment. Therefore, we encourage future researchers to evaluate the plausibility and effectiveness of the generated distractors to explore the significance of our proposed method thoroughly. Second, our current method required labeled responses to identify students' responses with incorrect answers. Scoring students' responses manually can be a very expensive and tedious procedure, especially in a large-scale assessment. However, as the current method attempts to extract students' misconceptions that could be located from their incorrect responses, it is necessary to score or use pre-labeled data set to properly implement the proposed method. This could somewhat limit the usability of the proposed method as locating domain specific and pre-labeled data can be a daunting challenge. However, we believe such limitations can be readily overcome by using automated essay scoring systems (see **Appendix C**) to generate labeled responses in advanced to implement the current method. Last, augmented intelligence approach of our method aim to create a systematic method to distractor development supporting content experts to make informed decisions using misconception clusters. Therefore, it is important to investigate whether content specialists, indeed, feel supported to make informed decisions in creating distractors. We encourage future research to carefully evaluate the affective factors of content experts in using this method to fully evaluate the capacity of the current method.

AUTHOR CONTRIBUTIONS

JS, QG, and MG contributed in conceptualization and formalization of research ideas of the study. JS located and organized the data. JS and QG performed the analysis. JS and MG wrote the first draft of the manuscript. All authors contributed to manuscript revision, read and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00825/full#supplementary-material

REFERENCES

Apaza, R. G., Cervantes, E. V., Quispe, L. C., and Luna, J. O. (2014). "Online courses recommendation based on LDA," in *Proceedings of the Symposium on Information Management and Big Data - SIMBig*, Peru, 42–48.

Arun, R., Suresh, V., Madhavan, C. V., and Murthy, M. N. (2010). "On finding the natural number of topics with latent dirichlet allocation: some observations," in *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining* (Berlin: Springer), 391–402. doi: 10.1007/978-3-642-13657-3_43

Ascalon, M. E., Meyers, L. S., Davis, B. W., and Smits, N. (2007). Distractor similarity and item-stem structure: effects on item difficulty. *Appl. Meas. Educ.* 20, 153–170. doi: 10.1080/08957340701301272

Barrios, F., López, F., Argerich, L., and Wachenchauzer, R. (2016). Variations of the similarity function of textrank for automated summarization. *arXiv* [Preprint]. arXiv:1602.03606

Bekkink, M. O., Donders, A. R., Kooloos, J. G., de Waal, R. M., and Ruiter, D. J. (2016). Uncovering students' misconceptions by assessment of their written questions. *BMC Med. Educ.* 16:221. doi: 10.1186/s12909-016-0739-5

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with PYTHON: Analyzing Text with the Natural Language Toolkit*. Newton, MA: O'Reilly Media, Inc.

Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM* 55, 77–84. doi: 10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022.

Briggs, D. C., Alonzo, A. C., Schwab, C., and Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice ite

Cao, J., Xia, T., Li, J., Zhang, Y., and Tang, S. (2008). A density-based method for adaptive LDA model selection. *Neurocomputing* 72, 1775–1781. doi: 10.1016/j.neucom.2008.06.011

Case, S. M., and Swanson, D. B. (2001). *Constructing Written Test Questions for the Basic and Clinical Sciences*, 2nd Edn. Philadelphia, PA: National Board of Medical Examiners.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. (2009). "Reading tea leaves: how humans interpret topic models," in *Proceeding of the 23rd Annual Conference on Neural Information Processing Systems*, (Iowa City: NIPC), 288–296.

Chingos, M. M. (2012). *Strength in Numbers: State Spending on K-12 Assessment Systems*. Available at: https://www.brookings.edu/research/strength-in-numbers-state-spending-on-k-12-assessment-systems/[accessed November 29, 2012]

Collins, J. (2006). Writing multiple-choice questions for continuing medical education activities and self-assessment modules. *Radiographics* 26, 543–551. doi: 10.1148/rg.262055145

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Appl. Psychol. Meas.* 33, 163–183. doi: 10.1177/0146621608320523

DiBattista, D., and Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *Can. J. Scholarsh. Teach. Learn.* 2:4. doi: 10.5206/cjsotl-rcacea.2011.2.4

Downing, S. M. (2006). "Selected-response item formats in test development," in *Handbook of Test Development*, eds T. M. Haladyna and S. M. Downing (Didcot: Taylor & Francis Group), 287–301.

Gierl, M. J., Bulut, O., Guo, Q., and Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: a comprehensive review. *Rev. Educ. Res.* 87, 1082–1116. doi: 10.3102/0034654317726529

Gierl, M. J., and Haladyna, T. M. (2013). *Automatic Item Generation: An Introduction. Automatic Item Generation: Theory and Practice*. New York, NY: Routledge.

Gierl, M. J., and Lai, H. (2013). Using automated processes to generate test items. *Educ. Meas.* 32, 36–50. doi: 10.1080/10401334.2016.1146608

Gierl, M. J., and Lai, H. (2016). "Automatic item generation," in *Handbook of Test Development*, 2nd Edn, eds S. Lane, M. Raymond, and T. Haladyna (New York, NY: Routledge), 410–429.

Griffiths, T. L., and Steyvers, M. (2004). Finding scientific topics. *Proc. Natl. Acad. Sci. U.S.A.* 101(Suppl. 1), 5228–5235. doi: 10.1073/pnas.0307752101

Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items*, 3rd Edn. Mahwah, NJ: Erlbaum.

Haladyna, T. M., and Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Appl. Meas. Educ.* 2, 37–50. doi: 10.1207/s15324818ame0201_4

Haladyna, T. M., and Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educ. Psychol. Meas.* 53, 999–1010. doi: 10.1177/0013164493053004013

Haladyna, T. M., and Rodriguez, M. C. (2013). *Developing and Validating Test Items*. New York, NY: Routledge. doi: 10.4324/9780203850381

Hoshino, Y. (2013). Relationship between types of distractor and difficulty of multiple-choice vocabulary tests in sentential context. *Lang. Test. Asia* 3, 1–14. doi: 10.1186/2229-0443-3-16

Irvine, S., and Kyllonen, P. (2002). *Item Generation for Test Development*. Mahwah, NJ: Erlbaum.

Konrad, M. (2017). *Topic Model Evaluation in Python with Tmtoolkit*. Available at: https://datascience.blog.wzb.eu/2017/11/09/topic-modeling-evaluation-in-python-with-tmtoolkit/ [accessed October 02, 2018].

Lai, H., Gierl, M. J., Touchie, C., Pugh, D., Boulais, A., and De Champlain, A. (2016). Using automatic item generation to improve the quality of MCQ distractors. *Teach. Learn. Med.* 28, 166–173. doi: 10.1080/10401334.2016.1146608

Meisner, R. (2018). Impact of semantic similarity to training responses on automated scoring accuracy. *Paper presented at the Annual Meeting of the National Council on Measurement in Education*, New York, NY.

Mihalcea, R., and Tarau, P. (2004). "Textrank: bringing order into text," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona.

Moreno, R., Martínez, R. J., and Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology* 2, 65–72. doi: 10.1027/1614-2241.2.2.65

Moreno, R., Martínez, R. J., and Muñiz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. *Psicothema* 27, 388–394. doi: 10.7334/psicothema2015.110

Moretti, A., McKnight, K., and Salleb-Aouissi, A. (2015). "Application of sentiment and topic analysis to teacher evaluation policy in the US," in *Proceedings of the 8th International Conference on Educational Data Mining*, Madrid, 628–629.

Mullis, I. V. S., Cotter, K. E., Fishbein, B. G., and Centurino, V. A. S. (2016). "Developing the TIMSS advanced 2015 achievement items," in *Methods and Procedures in TIMSS 2015*, eds M. O. Martin, I. V. S. Mullis, and M. Hooper (Chestnut Hill, MA: TIMSS & PIRLS), 1.1–1.17.

Nagata, R., Takamura, H., and Neubig, G. (2017). Adaptive spelling error correction models for learner english. *Procedia Comput. Sci.* 112, 474–483. doi: 10.1016/j.procs.2017.08.065

OECD (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: OECD Publishing. doi: 10.1787/9789264266490-en

Pan, Y. (2016). Heading toward artificial intelligence 2.0. *Engineering* 2, 409–413. doi: 10.1016/J.ENG.2016.04.018

Pennington, J., Socher, R., and Manning, C. (2014). "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, 1532–1543. doi: 10.3115/v1/D14-1162

Popenici, S. A., and Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Res. Pract. Technol. Enhanc. Learn.* 12:22. doi: 10.1186/s41039-017-0062-8

Popham, W. J. (2008). *Transformative Assessment*. Alexandria, VA: ASCD.

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., and Welling, M. (2008). "Fast collapsed gibbs sampling for latent dirichlet allocation," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (New York, NY: ACM), 569–577. doi: 10.1145/1401890.1401960

Rodriguez, M. C. (2011). "Item-writing practice and evidence," in *Handbook of Accessible Achievement Tests for all Student: Bridging the Gaps Between Research, Practice, and Policy*, eds S. N. Elliott, R. J. Kettler, P. A. Beddow, and A. Kurz (New York, NY: Springer), 201–216. doi: 10.1007/978-1-4419-9356-4_11

Rodriguez, M. C. (2016). "Selected-response item development," in *Handbook of Test Development*, 2nd Edn, eds S. Lane, M. Raymond, and T. Haladyna (New York, NY: Routledge), 259–273.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: competition, results, and future directions from a United States demonstration. *Assess. Writ.* 20, 53–76. doi: 10.1016/j.asw.2013.04.001

Shermis, M. D. (2015). Contrasting state-of-the-art in the machine scoring of short-form constructed responses. *Educ. Assess.* 20, 46–65. doi: 10.1080/10627197.2015.997617

Tarrant, M., Ware, J., and Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Med. Educ.* 9:1–8. doi: 10.1186/1472-6920-9-40 doi: 10.1186/1472-6920-9-40

Towns, M. H. (2014). Guide to developing high-quality, reliable, and valid multiple-choice assessments. *J. Chem. Educ.* 9:40. doi: 10.1021/ed500076x

Vacc, N. A., Loesch, L. C., and Lubik, R. E. (2001). "Writing multiple-choice test items," in *Assessment: Issues and Challenges for the Millennium*, eds G. R. Walz and J. C. Bleuer (Greensboro, NC: ERIC), 215–222.

Zheng, N. N., Liu, Z. Y., Ren, P. J., Ma, Y. Q., Chen, S. T., Yu, S. Y., et al. (2017). Hybrid-augmented intelligence: collaboration and cognition. *Front. Inform. Technol. Electron. Eng.* 18:153–179. doi: 10.1631/FITEE.1700053

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX A

Prompt—cell membrane item

List and describe three processes used by cells to control the movement of substances across the cell membrane.

Rubric for cell membrane

Key elements:

- Selective permeability is used by the cell membrane to allow certain substances to move across.
- Passive transport occurs when substances move from an area of higher concentration to an area of lower concentration.
- Osmosis is the diffusion of water across the cell membrane.
- Facilitated diffusion occurs when the membrane controls the pathway for a particle to enter or leave a cell.
- Active transport occurs when a cell uses energy to move a substance across the cell membrane, and/or a substance moves from an area of low to high concentration, or against the concentration gradient.
- Pumps are used to move charged particles like sodium and potassium ions through membranes using energy and carrier proteins.
- Membrane-assisted transport occurs when the membrane of the vesicle fuses with the cell membrane forcing large molecules out of the cell as in exocytosis.
- Membrane-assisted transport occurs when molecules are engulfed by the cell membrane as in endocytosis.
- Membrane-assisted transport occurs when vesicles are formed around large molecules as in phagocytosis.
- Membrane-assisted transport occurs when vesicles are formed around liquid droplets as in pinocytosis.
- Protein channels or channel proteins allow for the movement of specific molecules or substances into or out of the cell.

Rubric:
3 points
Three key elements
2 points
Two key elements
1 point
One key element
0 points
Other.

APPENDIX B

Representative key words of topic clusters

Topic	Key words	Summary
1	Cell, osmosis, water, diffusion, membrane, process, permeable, moving	Three processes used by cells to control the movement of substances across the cell membrane are being selectively or semi permeable, osmosis, and diffusion
2	Transport, active, diffusion, passive, osmosis, processes, facilitated, type	Three types of controlled movement of substances across the cell membrane include passive transport, active transport, and diffusion
3	Cell, substance, membrane, way, moves, cytoplasm, goes, organism	Another one is where the organism extends out sections of its cell membrane and fills it with cytoplasm while the opposite end goes away and it moves by a crawling type movement
4	Cells, blood, body, make, flow, need, brain, send	The movements of substances across the cell membrane flow through the blood streams
5	Cell, membrane, wall, help, nucleus, things, outside, inside	Three processes used by cells to control the movement of substances across the cell membrane is flagella which helps the cell get through the membrane, the nucleus that is the control center, and the cell wall to protect the cell from any unwanted cells or anything unwanted
6	Cell, waste, food, gets, stuff, nutrients, needed, needs	The Golgi bodies help by getting rid of stuff not needed in the cell
7	Protein, proteins, cell, enzymes, synthesis, channel	The cell uses three basic processes for movement across the membrane one is the flagellum, another cytoplasm and finally the protein in the ribonuclease acid
8	Cell, membrane, movement, control, substances, helps, plasma, different	Pores in the membrane allow substances in and out of the cell and Golgi body helps transport substances in and out of cell
9	Cells, use, proteins, way, membrane, ribosomes, carry, proteins	Cells use vesicles, transport chains, and proteins to control the movement of substances across the cell membrane
10	Cell, substance, membrane, diffusion, concentration, substances, movement, uses	Osmosis is the movement of water going from a low concentration to a high concentration in the cell membrane
11	Golgi, nucleus, proteins, apparatus, ribosomes, reticulum, endoplasmic, use	The ribosomes produce the energy for the cell the Golgi apparatus gets rid of waste and the nucleus hold all the information and DNA
12	Cell, things, wall, membrane, inside, getting, substances, lets	Cell wall makes the plant cell stiff but also keeps out unwanted items or organism's cell membrane lets things in and out of the cell with permission from the nucleus and chloroplast help the plants maintain energy
13	Like, flagellum, flagella, use, cilia, cell, helps, help	One way of movement is the use of flagellum which is a long tail like structure the moves behind the cell
14	Cells, substances, use, process, place, organelles, moving, help	Another processes but which cells use to control the substances that cross the cell membranes are the phospholipids that line that cell wall these help keep unwanted thing out as well
15	Movement, cells, control, used, cell, processes, substances, membrane	Three processes that cells use to control the movement of substances across the cell membrane is protein synthesis, transfusion, and moving waste out
16	Cells, mitosis, meiosis process, reproduce, make, makes, meiosis	Another processes but which cells use to control the substances that cross the cell membranes are the phospholipids that line that cell wall these help keep unwanted thing out as well
17	Cell, controls, membrane, nucleus, goes, wall, tells, comes	The cell uses three processes by the names of meiosis, mitosis, and cell reproduction
18	Cell, uses, energy, things, membrane, moves, mitochondria, endocytosis	The nucleus controls everything and the mitochondria tell what enters and leaves the cell
19	Respiration, cellular, reproduction, photosynthesis, process, food, division, homeostasis	Endocytosis which is part of active transport, where the cell uses energy to pull items through the selectively permeable membrane
20	mRNA, tRNA, RNA, DNA, information, translation, transcription, messages	It does this to maintain homeostasis the cell moves oxygen in and carbon dioxide of out of the cell through the process of cellular respiration
21	Cell, membrane, certain, let, things, substances, enter, allow	mRNA carries messages from the nucleus to other organs tRNA transports DNA to places with in the cell rRNA
22	Anaphase, telophase, thing, prophase, metaphase, second, interphase, know	Three of the processes that cells use to control movement into and out of the cell membrane are protein channels that let substances pass through them, endosymbiosis allows large substances to enter, and exocytosis allow larger substances to exit the cell

ORIGINAL RESEARCH
published: 26 April 2019
doi: 10.3389/fpsyg.2019.00853

The Expanded Evidence-Centered Design (e-ECD) for Learning and Assessment Systems: A Framework for Incorporating Learning Goals and Processes Within Assessment Design

Meirav Arieli-Attali[1,2]*, *Sue Ward*[3], *Jay Thomas*[3], *Benjamin Deonovic*[1] *and Alina A. von Davier*[1]

[1]ACTNext, ACT Inc., Iowa City, IA, United States, [2]Fordham University, New York City, NY, United States, [3]ACT Inc., Iowa City, IA, United States

OPEN ACCESS

Edited by:
*Frank Goldhammer,
German Institute for International Educational Research (LG), Germany*

Reviewed by:
*Russell G. Almond,
Florida State University, United States
Gabriels Nagy,
Christian-Albrechts-Universität zu Kiel, Germany*

***Correspondence:**
*Meirav Arieli-Attali
meirav.attali@act.org*

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *10 December 2018*
Accepted: *01 April 2019*
Published: *26 April 2019*

Citation:
Arieli-Attali M, Ward S, Thomas J, Deonovic B and von Davier AA (2019) The Expanded Evidence-Centered Design (e-ECD) for Learning and Assessment Systems: A Framework for Incorporating Learning Goals and Processes Within Assessment Design. Front. Psychol. 10:853. doi: 10.3389/fpsyg.2019.00853

Evidence-centered design (ECD) is a framework for the design and development of assessments that ensures consideration and collection of validity evidence from the onset of the test design. Blending learning and assessment requires integrating aspects of learning at the same level of rigor as aspects of testing. In this paper, we describe an expansion to the ECD framework (termed e-ECD) such that it includes the specifications of the relevant aspects of learning at each of the three core models in the ECD, as well as making room for specifying the relationship between learning and assessment within the system. The framework proposed here does not assume a specific learning theory or particular learning goals, rather it allows for their inclusion within an assessment framework, such that they can be articulated by researchers or assessment developers that wish to focus on learning.

Keywords: task design, technology-based assessment, blended assessment and learning, development framework, Evidence model

INTRODUCTION

There is a growing need for the development of assessments that are connected and relevant to learning and teaching, and several attempts have been made in recent years to focus on this topic in conferences and journals. For example, Mark Wilson's 2016 June and September presidential messages in the National Council for Measurement in Education's newsletter addressed Classroom Assessment, and this topic was also the conference theme for the following 2 years, 2017 and 2018. The journal *Assessment in Education: Principles, Policy & Practice* recently devoted a special issue on the link between assessment and learning (volume 24, issue 3, 2017). The issue focused on the developments in the two disciplines which, despite mutual influences, have taken distinctly separate paths over time. In recent years, systems that blend learning and assessment have been proposed all over the world (e.g., Razzaq et al., 2005; Shute et al., 2008; Feng et al., 2009b; Attali and Arieli-Attali, 2014; Straatemeier, 2014). While within the educational measurement field, there are established standards and frameworks for the development of reliable and valid assessments, those rarely take learning aspects into account. As part of our own effort to develop a blended learning and assessment system, we identified a need for a formal framework of development that includes aspects of learning at the same level of detail and rigor as aspects of testing. This paper describes our general approach at expanding an assessment framework, with some examples from our system to better illustrate the abstract concepts.

Our approach at expanding a principled assessment design is primarily concerned with the inclusion of three dimensions: *aspects of learning*, such as the ability to incorporate the change over time in the skills to be measured at the conceptual level; *aspects of interactive and digital instructional content*, such as simulations, games, practice items, feedback, scaffolds, videos, and their associated affordances for the data collection in rich logfiles; and *measurement models for learning* that synthesize the complexities of the digital instruction and data features.

The expanded framework proposed here allows for the design of systems for learning that are principled, valid, and focused on the learner. Systems designed in this framework are intrinsically connected with the assessment of the skills over the time of instruction, as well as at the end, as summative tests, if so desired. This type of systems has an embedded efficacy structure, so that additional tests can be incorporated within. Learning and assessment developers, as well as researchers, can benefit from such a framework, as it requires articulating both the assessment and learning intended goals at the start of the development process, and it then guides the process to ensure validity of the end-product. The framework proposed here does not assume a specific learning theory or particular learning goals, rather it allows for their inclusion within the assessment framework. The measurement perspective, combined with the learning sciences perspective in the development of content, provides a new and significant shift in the modern development of leaning and assessment systems.

We chose to expand the well-known evidence-centered design framework (ECD; Mislevy et al., 1999, 2003, 2006). The ECD formulates the process of test development to ensure consideration and collection of validity evidence from the onset of the test design. The ECD is built on the premise that a test is a measurement instrument with which specific claims about the test scores are associated, and that a good test is a good match of the test items and the test takers' skills. The ECD framework defines several interconnected models, three of which form the core of the framework and are relevant to our discussion: the Student model(s), Evidence model(s), and Task model(s) (the combination of the three models is also called the Conceptual Assessment Framework; CAF; see **Figure 1**). Note that in more recent publications of the ECD, the Student model is termed a Proficiency model (e.g., Almond et al., 2015).

The Student or the Proficiency model(s) specifies the knowledge, skills, and ability (KSA; which are *latent* competencies) that are the target of the test. This model can be as simple as defining one skill (e.g., the ability θ) or a map of interconnected subskills (e.g., fractions addition, subtractions, multiplication, and division are interconnected subskills that form the map of knowing fractions). The latent competencies that are articulated and defined in this model establish the conceptual basis of the system, and they are often based on a theory or previous findings related to the goal of the assessment.

Since we cannot tap directly into the latent competencies, we need to design tasks/test items such that they will elicit behaviors that can reflect on or indicate about the latent competencies. This is the role of the Task model(s). The Task model specifies the *tasks features* that are supposed to elicit the observables, and only them, such that to allow inferences about the latent competencies. For example, if the assessment is intended to measure "knowledge of operating with fractions,"

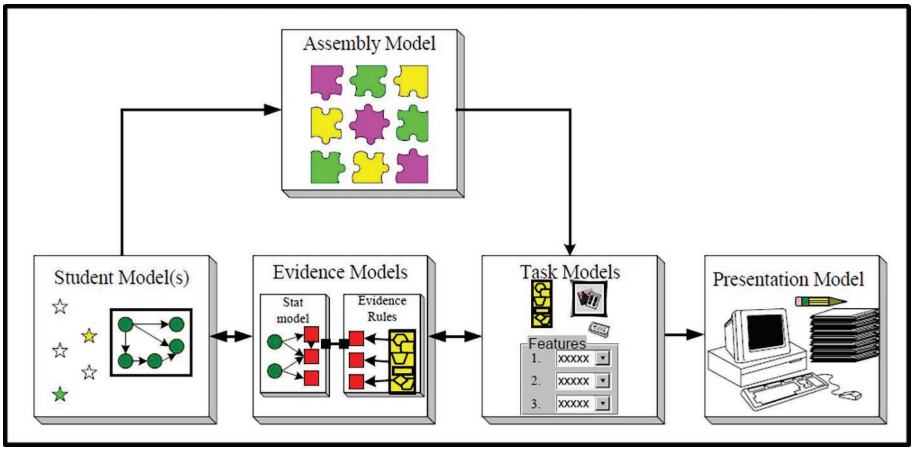


FIGURE 1 | The core models within the ECD framework (from Mislevy Almond & Lucas, © 2003 Educational Testing Service; used with permission); note that later versions term the Student model as Proficiency model.

the tasks should be designed with care such that reading ability is not an obstacle to perform well on the task and express one's fractions knowledge.

The Evidence models then make the connection between the latent competencies [specified by the Student/Proficiency model(s)] and the observables [behaviors elicited by the Task model(s)]. In other words, the Evidence models are the connecting link. The Evidence models include the measurement model, comprised of the rubrics, the scoring method, and the statistical method for obtaining a total score(s). See **Figure 1** for a diagram of the ECD and specifically the three CAF models (note that latent competencies are symbolized as circles, while observables as squares; and the connection between the circles and squares are shown in the Evidence models).

Two important additional models are the Assembly model and the Presentation model (see **Figure 1**). The Assembly model defines how the three models in the CAF (the Student/Proficiency, Task, and Evidence models) work together and specifically determines the conditions for reliability and validity of the system. As part of the Assembly model, the developers determine the number of items/tasks and their mix ("constraints") such they provide the necessary evidence and are balanced to properly reflect the breadth and diversity of the domain being assessed. The Presentation models are concerned with different ways to present the assessment, whether it is a paper-and-pencil test, a computer-based test, a hands-on activity, etc. We will elaborate on and delve deeper into each of the models as part of the expansion description below; for more details on the original ECD, see Mislevy et al. (2003, 2006).

There are other alternatives frameworks for the design and development of assessment that follow a principled approach, such as the Cognitive Design System (Embretson, 1998), the Assessment Engineering framework (Luecht, 2013), the Principled Design for Efficacy framework (Nichols et al., 2015), or the Principled Assessment Design framework (Nichols et al., 2016). These frameworks may be perceived as alternatives to the ECD, and one might find any of them as a candidate for a similar expansion the way we demonstrate executing for the ECD in this paper. The reason there were several assessment frameworks developed over the years stem from the need to ensure validity of assessment tools. Although traditional assessments were developed for about half a century without a principled approach (i.e., by following an assessment manual and specifications) and validity was verified after development, the advantage of following a principled framework such as the ECD or others is particularly evident when the goal is to assess *complex competencies* (e.g., problem solving, reasoning, collaborative work) and/or when using *complex performance tasks* (e.g., multidimensional tasks such as performance assessment, simulations or games on computer or otherwise). In these cases, it is important to explicitly identify the relevant competencies and behaviors and how they are connected, because the complexity of the focal competencies and/or the rich data that the tasks provide might pose difficulties in making inferences from behaviors to competencies. ECD has been also successfully applied to address the challenges of simulation- and game-based assessment (Rupp et al., 2010a; Mislevy, 2013; Kim et al., 2016).

MOTIVATION FOR A PRINCIPLED APPROACH TO THE DESIGN AND DEVELOPMENT OF A LEARNING AND ASSESSMENT SYSTEM

Learning and assessment, although both relate to the process of determining whether or not a student has a particular knowledge, skill, or ability (KSA), differ substantially in the way they treat KSAs. The main difference between an assessment tool and a learning tool is in the *assumption* about the focal KSA, whether it is fixed or dynamic at the time of interacting with the tool. The Student/Proficiency model in the ECD describes a map of competencies (KSAs), and as in most psychometric models for testing, the assumption is of a latent trait, which is "fixed" at the time of taking the test. The purpose of an assessment is thus to "detect" or "diagnose" that fixed latent KSA at a certain point in time, similar to any measurement tool (e.g., a scale measuring a person's weight at a particular point in time). On the other hand, the main purpose of a learning tool, such as a computer tutoring system, is to "move" the learner from one state of knowledge to another – that is, the concern is first and foremost with the *change* in KSAs over time, or the *transition*. Of course, an assessment tool *per se* cannot drive the desired change unless deliberate efforts are implemented in the design of the system (similar to a scale which will not help with weight loss unless other actions are taken). Thus, systems that aim at blending assessment and learning cannot implement ECD as is, since ECD is inherently a framework to develop assessments and not learning.

Moreover, the availability of rich data collected *via* technology-enhanced learning and assessment systems (e.g., trial and error as part of the learning process, hint usage) poses challenges, as well as promises, for assessment design and the decision process of which actions to allow and what to record, either to promote efficient learning or to enable the reliable assessment of the learning in order to make valid inferences about KSAs. Computational Psychometrics (von Davier, 2017), an emerging discipline, blends theory-based methods and data-driven algorithms (e.g., data mining and machine learning) for measuring latent KSAs. Computational Psychometrics is a framework for analyzing large and often unstructured data, collected during the learning or performance process, on a theoretical learning and psychometric basis. We also combine aspects of Computational Psychometrics in our expanded design framework, similar to previous accounts that integrated data mining into ECD (e.g., Mislevy et al., 2012; Ventura and Shute, 2013). Combining data-driven algorithms into ECD allows knowledge discovery and models' update from data, thereby informing the theory-based Student/Proficiency model and enriching the Evidence model.

Attempts to develop innovative assessments within games or as part of complex skills assessment and learning also brought about variations or expansions to ECD (e.g., Feng et al., 2009a; Conrad et al., 2014; Grover et al., 2017). One characteristic of ECD variants focuses on the task and its connection to the Evidence model. Since game-play and the rich data from complex assessments often result in sequences of actions, not all of which are relevant to the target competencies, researchers may follow an ECD approach with expansion with respect to the action-data, to specify which actions are relevant and should be included in the Evidence model and in what way (i.e., expansion on the scoring rules or both scoring and Task model). Such an attempt was done by Grover et al. (2017). Grover and her colleagues expanded on the scoring rules by employing data driven techniques (e.g., clustering, pattern recognition) in addition to theory-based hypotheses, to guide the definition of the scoring rules. Another interesting variation is the experiment-centered design by Conrad et al. (2014), which illustrated an expansion on the scoring and the Task model. This approach uses an ECD-like process to simultaneously encode actions of players in one way for game design and another way for assessment design. Because the game design dictates feedback on actions, and subsequent game options may depend on student's actions, the game designer needs to encode the actions differently than a researcher or an assessment designer, who is primarily interested in estimating whether a student possesses the focal skill. In this procedure, the model is first postulated around the task (experiment), and then applied separately as two models (versions), one for the game designer, and one for the researcher, each focused on a different encoding of student actions. However, there is only one Evidence model for inferring KSAs, derived from the researcher's version of the task encoding (the assessment variant scoring rule). In this way, the adaptation of the ECD allowed adding the assessment as a "layer" on top of the game design (stealth assessment), while ensuring coordination between these two layers.

Work by Feng et al. (2009a) is particularly relevant in this context. The authors examine an adaptation of the ECD for learning data (ECDL), applied retroactively to the ASSISTments data (Heffernan and Heffernan, 2014). The ECDL is an ECD with an augmented *pedagogical model*, which has links to all three models of the CAF (Proficiency, Evidence, and Task). The pedagogical model refers to the learning and learners' characteristics, including learning effectiveness and efficiency (e.g., reducing cognitive load, increasing difficulty gradually during presentation, adapting the presentation of content, and decomposing multistep problems to sub-steps), as well as learner engagement factors. Since ASSISTments was initially developed without ECD in mind, the analysis retroactively checks which claims can support a validity argument that an item with its hints and scaffolds serves the learning goal. This is done by identifying (within each item) the KSAs required to answer it correctly, tagging each as "focal" or "unfocal." The focal KSAs are the ones which the hints/scaffolds should address. The relation between the focal and unfocal also serves as an indication of the system's efficacy [a system with a high proportion of unfocal KSAs is less efficient than a system with a low proportion, because this reflects the proportion of KSAs not taught (scaffolded)]. In sum, Feng and his colleagues demonstrated how an existing learning product can be analyzed (and potentially improved) using an ECDL framework.

Common to the various adaptations of ECD is that they were task driven. First came the tasks; then came the ECD analysis, which resulted in adapting the ECD to address the complexity and intuition that were built into the tasks, expressed as an expansion on one of the three models in the CAF. While in the first two examples of Conrad et al. (2014) and Grover et al. (2017), the revised ECD focused on how to encode the task data to feed into the Evidence model, Feng et al.'s (2009a) study goes further, suggesting a pedagogical model that is feeding and being fed by all three CAF models – Proficiency, Evidence, and Task. However, this pedagogical model seems somewhat like a "black box" that retroactively includes the intuitions that specified the product design (e.g., how hints and scaffolds were determined). Additionally, it neither specifies the nature of the connections with the original ECD models nor does it inform how to design a learning product from scratch (i.e., a principled approach to development).

We offer a comprehensive expansion of the ECD framework, such that learning aspects are specified for each of the three models in the CAF and are determined *a priori* to the system design. We describe the expanded full CAF first, followed by a focus on each expanded model with examples. We then discuss the Assembly model, which allows for the specification of the relationship between assessment and learning. We conclude with ramifications of the expanded framework for the development of *adaptive* systems. We include examples to better illustrate the general ideas, along with directions for alternative decisions, to emphasis the generalizability of the expanded framework.

THE EXPANDED ECD MODEL

In our expanded ECD framework (e-ECD), we find it necessary to expand on all three Student/Proficiency, Evidence, and Task models. We do so by adding a *learning layer*, in parallel to the assessment layer. This learning layer can be viewed as a breakdown of a pedagogical model (Feng et al., 2009a) to three components, the conceptual (student/proficiency), behavioral (task), and statistical (evidence) components. Thus, each original ECD model now has an additional paired learning model, culminating in six models. We call each assessment-learning pair an expanded model (e-model), i.e., the e-Proficiency model, the e-Task model, and the e-Evidence model (see **Figure 2**). Note that we refer to the original Proficiency model as the KSA model (Knowledge, Skills, and Ability), which is now part of the e-Proficiency model.

Within each e-model, we denote an "observational" layer for the assessment aspect (these are the original ECD models with slight title change; the KSA model, Task model, and Observational-Evidence model) and a "transitional" layer for the learning aspect (these are the new models that address learning). The three new learning models include the following: (1) at the conceptual

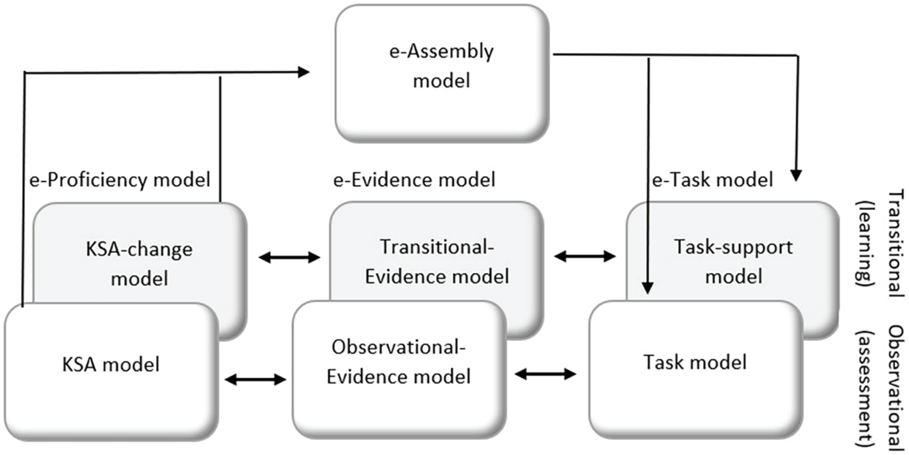


FIGURE 2 | Expanded ECD (e-ECD) for learning and assessment systems.

latent level and part of the e-Proficiency model – the transitional layer specifies *learning processes* as the latent competency that the system targets. We denote it as the KSA-change model; (2) at the behavioral level and part of the e-Task model – the transitional layer specifies principles and features of *learning support* that guides the design of tasks (customized feedback, scaffolds, hints, solved examples, solution, or guidance to digital instructional content such as animation, simulation, games, and videos). We denote it as the Task-support model; and (3) at the statistical level and part of the e-Evidence model – the transitional layer specifies the links between the learner's support usage and the target learning processes, to allow inferring from behaviors to latent learning (e.g., the efficiency of the support used in achieving learning). The data could be large process data and may reveal behavior patterns that were not identified by the human expert in the original e-Proficiency model. In this framework, the e-Proficiency model and the e-Evidence model are supposed to "learn" in real time (be updated) with the new knowledge inferred from the data. We denote it as the Transitional-Evidence model.

We include also an expansion on the Assembly model, denoted e-Assembly model. In addition to determining the number and mix of tasks, the e-Assembly model also includes the specification about the relationship between the assessment component and the learning component of the system and determines how they all work together. In other words, the assembly model determines the "structure" of the system, e.g., when and how learning materials appear and when and how assessment materials appear, and the rules for switching between the two.

Consider the following situation: a student is using a system for learning and assessment to learn and practice scientific reasoning skills. At some point, the student gets an item wrong. In a regular assessment system, another item will follow (often without any feedback about the correctness of the response) – and if the system is an adaptive testing system, the student will receive an easier item, but not necessarily with the same content as the item with the incorrect response. In a blended learning and assessment system, the approach is different. *Detecting a "weakness" in knowledge is a trigger to foster learning.* How should the system aim at facilitating learning? There are several different options, from providing customized feedback and hints on how to answer that specific item, presenting scaffolds for the steps required or eliciting prior knowledge that is needed to answer that item, addressing specific misconceptions that are known to be prevalent for that specific node of KSA, up to re-teaching the topic and showing worked examples, and/or presenting similar items to practice the skill. In many learning products today, this process of defining the learning options is conducted using content experts according to implicit or explicit learning goals. Using a principled approach to development will dictate that the definition of the options for learning should be *explicitly* articulated at the level of the Task-support model, and these features are to be in line with the explicit conceptual learning/pedagogical model that describes how to make that shift in knowledge, i.e., the KSA-change model. The links between the supports and the conceptual KSA-change are defined in the Transitional-Evidence model *via* statistical models, which provide the validity learning argument for the system.

In the development of an assessment system that blends learning, we wish to help students learn, and to validate the claim that learning occurred, or that the system indeed helped with the learning *as intended*. The KSA-change specifies the type of changes (learning/transitions) the system is targeting, and based on that, the tasks and the task supports are defined. In other words, the first step is to define the "learning shifts" or how to "move" in the KSA model from one level/node to the next. The next step is to define the observables that need to be elicited and the connections between the learning shifts and the observables. We elaborate on each of the expanded models below.

Our expanded framework shows how to incorporate a learning theory or learning principles into the ECD and can be applied using different learning approaches. We illustrate this process by using examples from Knowledge-Learning-Instruction

(Koedinger et al., 2012) among others, but this process can be applied using other learning approaches (and we provide some directions).

Expanded Proficiency Model

In the ECD framework, the Student/Proficiency model defines the Knowledge, Skills, and Ability (KSA) that the assessment is targeting. Although in early publications of the ECD, it is called a Student model, in recent contexts, it is called a "Proficiency model" (e.g., Feng et al., 2009a; Almond et al., 2015), or referred to as a "Competency model" (e.g., Arieli-Attali and Cayton-Hodges, 2014; Kim et al., 2016), and it can also be perceived as a "Construct map" (Wilson, 2009). A similar notion in the field of Intelligence Tutoring Systems is a "Domain model" (Quintana et al., 2000), a "Knowledge model" (Koedinger et al., 2012; Pelánek, 2017), or a "Cognitive model" (Anderson et al., 1995). In the Intelligence Tutoring Systems' literature, the term "Student model" is reserved to a specific map of skills as estimated for a *particular student* – which is an overlay on the domain model (aka the expert model). Within ECD, the Student/Proficiency model includes both the desired skills (that an expert would possess) and the updated level of skills for each particular student following responses on assessment items. To avoid confusion, within our expanded ECD, we refer to it by the general name of a KSA model.

The KSAs are assumed to be latent, and the goal of the assessment is to infer about them from examinee's responses to test items. When the assessment tool is also intended to facilitate learning (i.e., the system provides supports when the student does not know the correct answer), the assumption is that the student's level of KSA is *changing* (presumably becoming higher as a result of learning). In the e-ECD, we define a "KSA-change model" that together with the original KSA model creates the expanded-Proficiency model (e-Proficiency model). The KSA-change model specifies the latent *learning processes* that need to occur in order to achieve specific nodes in the KSA model. Each node in the KSA model should have a corresponding *learning-model* in the KSA-change model, which may include prerequisite knowledge and misconceptions, and/or a progression of skills leading up to that KSA node, with the pedagogical knowledge of how to make the required knowledge-shift. Some examples of learning models are learning progressions (National Research Council (NRC), 2007; e.g., Arieli-Attali et al., 2012) a Dynamic Learning Map (Kingston et al., 2017), or learning models based on the body of work on Pedagogical Content Knowledge (Posner et al., 1982; Koehler and Mishra, 2009; Furtak et al., 2012). The importance of Pedagogical Content Knowledge is in considering the interactions of *content information*, *pedagogy*, and *learning theory*. Another approach from the learning sciences and artificial intelligence is the Knowledge-Learning-Instruction framework (KLI; Koedinger et al., 2012), which provides a taxonomy to connect knowledge components, learning processes, and teaching options. We will illustrate our KSA-change model specification using the KLI framework, but we will define the e-Proficiency model in a general way such that any other learning theory can be applied instead.

Specifying and explicitly articulating the latent learning processes and progressions that are the target of the learning is a crucial step, since this is what will guide the specification of both the e-Task model and the e-Evidence model. In the following sections, we elaborate and illustrate the KSA and KSA-change models that constitute the e-Proficiency Model.

The Assessment Layer of the e-Proficiency Model – The KSA Model

A KSA model includes *variables* that are the features or attributes of competence that the assessment is targeting. The number of variables and their grain size are determined by the potential use of the assessment, and it can range from 1 (e.g., the θ in college admission tests such as the GRE, SAT, and ACT) to several subskills arranged in a map or a net (e.g., a net example, see Mislevy et al., 1999; a math competency map, see Arieli-Attali and Cayton-Hodges, 2014; two versions of a game-based physics competency model, see Kim et al., 2016). These variables can be derived by conducting a *cognitive task analysis* of the skill by experts, analyzing the content domain, or relying on a theory of knowledge and research findings. The variables and their interconnections create a map in which each variable is a *node* connected by a *link* with other nodes (variables). Following analysis of data from student responses (and using the statistical models), values on these variables define the level of mastery or the probability that a particular student possess those particular sub-skills (nodes), i.e., a value will be attached to each node.

As part of our development of a learning and assessment system, called the Holistic Educational Resources & Assessment (HERA) system for scientific thinking skills, we developed a KSA model for *data interpretation* skill. **Figure 3** depicts part of the model. Specifically, we distinguish three main skills of data interpretation depending on the data representation (*Table Reading*, *Graph Reading*, and the skill of interpreting data from *both tables and graphs*), and each skill is then divided to several subskills. For example, in *Table Reading* skill, we distinguish between *locating data points*, *manipulating data*, *identifying trend*, and *interpolation and extrapolation*. Note that these same subskills (albeit in a different order) appear also under *Graph Reading* skill, but they entail different cognitive ability. The skill of *Tables and Graphs* includes *comparing*, *combining*, and *translating* information from two or more different representations.

Although KSA models often specify the links between nodes, and may even order the skills in a semi-progression (from basic to more sophisticated skills) as in the example of the HERA model in **Figure 3**, a knowledge model often does not specify *how to move* from one node to the next, nor does it explicitly define learning processes. To that end we add the learning layer in the e-Proficiency model – the KSA-change model.

The Learning Layer in the e-Proficiency Model – The KSA-Change Model

Defining a learning layer within the e-Proficiency model makes room for explicit articulation of the learning processes targeted by the learning and assessment system. The idea is for these

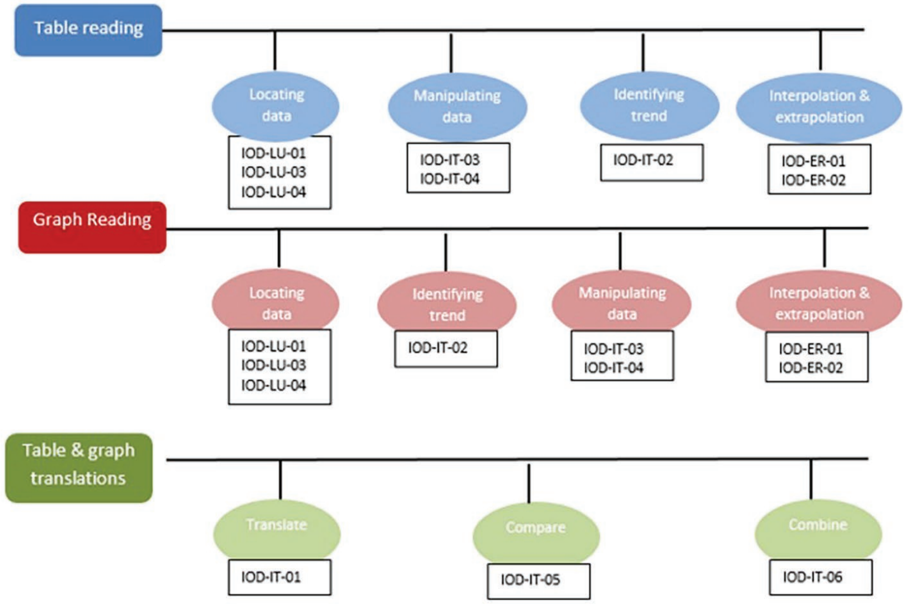


FIGURE 3 | The KSA model for the HERA system for scientific reasoning skills.

specifications to be the result of purposeful planning, rather than a coincidental outcome of system creation. In the Intelligence Tutoring literature, developers consider what they call the "Learner model" (Pelánek, 2017) or the "Educational model" (Quintana et al., 2000) or more generally, processes for knowledge acquisition (Koedinger et al., 2012). This model can also be viewed as the "pedagogical model" and apply principles of Pedagogical Content Knowledge (Koehler and Mishra, 2009; Furtak et al., 2012). We call this model the "KSA-change Model" for generalizability and to keep the connection with the original KSA model, with the emphasis on the *change* in KSA. Using the title "change" makes room also for negative change (aka "forgetting"), which albeit not desirable, is possible.

A KSA-change model is the place to incorporate the specific learning theory or learning principles (or goals) that are at the basis of the systems. Similar to the way a KSA map is created, the KSA-change map should specify the learning aspects of the particular skills. Here we provide a general outline for how to specify a KSA-change model, but in each system this process may take a different shape.

A KSA-change model may include variables of two types:

1. Sequences of knowledge components, features or attributes
2. Learning processes within each sequence

These two types of variables define the learning *sequences* and *processes* that are needed to facilitate learning. The KSA-change variables are derived directly from the KSA model, such that each node/skill in the KSA model has a reference in the KSA-change model in the form of how to "move" students to learn that skill.

Given a specific skill (node in the map), this may be done in two stages: (1) the first step is to define the (linear) *sequence* of pre-requisites or precursors needed to learn that target skill (node). For example, Kingston and his colleagues (Kingston et al., 2017) developed Dynamic Learning Maps in which each of the target competencies are preceded with three levels of precursor pieces of knowledge (initial precursor, distal precursor, and proximal precursor) and succeeded by a successor piece of knowledge, together creating what they called "Linkage levels." When defining the sequence of precursors attention should be given to the grain size, as well as to specific features or attributes of these precursors. In KLI terminology (Koedinger et al., 2012), this would mean to characterize the *Knowledge Components* of the subskills. Some Knowledge Components are: fact, association, category, concept, rule, principle, plan, schema, model, production; and whether it is verbal or non-verbal, declarative or procedural; or integrative knowledge (2) the second step is to characterize the learning sequence by which kind of learning *process* is required to achieve the learning. For example, applying the KLI taxonomy (Koedinger et al., 2012), we can assign to each precursor (knowledge component) a specific learning process that is presumed to make the desired knowledge shift. The KLI framework characterizes three kinds of learning processes: *memory and fluency building*, *induction and refinement*, and *understanding and sense-making*. Specifying which kind of process is needed in the particular learning sequence is necessary for subsequent decisions about the supports to be provided. For example, if the focal learning process is *fluency building*, this implies that the learning system should provide practice opportunities for that KSA. In contrast, if the focal learning process for a different KSA is *understanding and sense making*, then the learning system should provide explanations and examples. **Figure 4** illustrates a general e-Proficiency model with an artificial example of adding-on the learning processes to a knowledge sequence built off of three prerequisites and a successor piece.

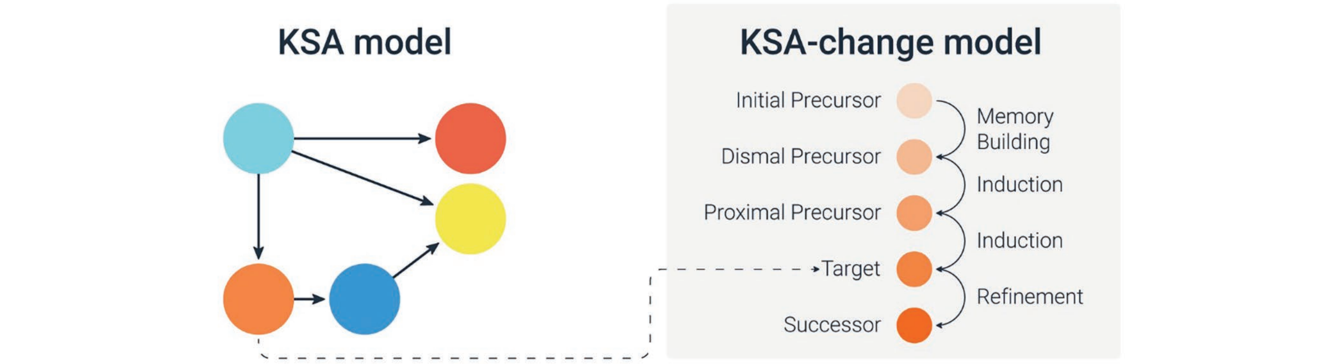


FIGURE 4 | A general diagram of the e-Proficiency model (the orange node in the KSA model is specified in the KSA-change model for learning sequence and learning processes). Similarly, we can construct a sequence for each of the other nodes (the blue, pink, and red nodes).

Applying the above approach to the HERA learning and assessment system, let us focus on the subskill of *interpolation and extrapolation from data in a graph* (the last red circle in the progression of *Graph Reading* skill in **Figure 3**). Based on our guidelines above, the first step would be to determine a sequence of subskills/precursors and to characterize them, and then as a second step to specify the cognitive process(es) that would make the transition from one subskill to the next. **Figure 5** presents one section of the KSA-change of the HERA system for the subskill of *interpolation and extrapolation in a graph*. The model specifies the proximal, distal, and initial precursors as follows: the proximal precursor = *identifying the rate of change in the dependent variable (y-variable) as the independent variable (x-variable) changes*; distal precursor = *being able to locate the y-value for a certain x-value point on a graph, and find adjacent points and compare the relative values*; initial precursor = *understanding that the two variables in a graph are co-related.* Now applying the KLI knowledge components characterization, the proximal precursor (identifying rate of change) may be characterized as "rule"; the distal precursor (locate points and compare) as "schema"; and the initial precursor (two variables are co-related) as a "concept."

Next, we determine the cognitive processes that foster the transition from one subskill to the next. For example, given an understanding of the co-variation of x and y (the initial subskill) students need to practice finding the y-points for different x-points to create the mental schema and build *fluency* with locating points and particularly two adjacent points. However, to "jump" to the next step of identifying the trend and the rate of change requires *induction and refinement* to derive the rule. The last transition from identifying rate of change to perform interpolation & extrapolation requires *sense making and deduction* – deducing from the rule to the new situation. Given the specific learning processes, we can later define which learning supports would be most appropriate (e.g., practice for fluency building, worked example and comparisons for induction, and explanation for sense making and deduction). The model in **Figure 5** shows the different learning processes as the transitions (arrows) required between the subskills in the sequence. This is the learning model for the specific skill in focus, and is usually derived based on expert analysis. The model in **Figure 5** also specifies particular misconceptions that students often exhibit at each level. Specifying misconceptions may also help determine which feedback and/or learning aid to provide to students. We show in the next section how to define Task and Task-support models based on this example.

There are several decisions that are taken as part of the model specifications. One of them is the grain-size of each precursor. An alternative KSA-change model can be determined with smaller or larger grain size subskills. Another decision is whether to adopt a three-level precursor skill structure, or alternatively focus on only one precursor and the different misconceptions students may have. Researchers and developers are encouraged to try different approaches.

We propose to derive the KSA-change variables by conducting a *learning process analysis* by experts, i.e., an analysis of the pedagogical practices in the content domain or relying on a theory of learning in that domain, similar to the way we illustrated above (by using the KLI taxonomy). This is also parallel to the way a KSA model is derived based on *cognitive task analysis* or domain analysis. The KSA-change model constitutes a *collection* of sequences (and their processes), each addressing one node in the KSA model (as illustrated in **Figures 4, 5**). This can also be viewed as a two-dimensional map, with the sequences as the second dimension for each node.

Similar to updating the KSA model for a student, here too, following analysis of data from student responses and student behaviors in using the learning supports, values on the KSA-change variables indicate level or probability that a particular student has gone through a particular learning process (or that a particular knowledge shift was due to the learning support used). We will discuss this in more detail in the e-Evidence model section.

Expanded Task Model

In the original ECD framework, the Task model specifies the features of tasks that are presumed to elicit observables to allow inference on the target KSA. An important distinction

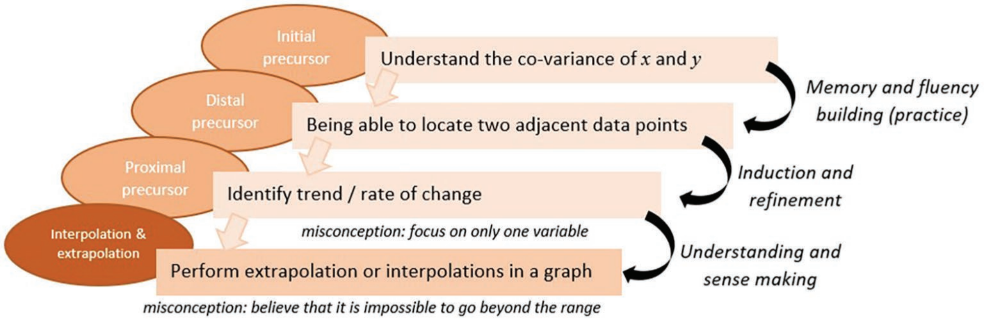

Initial precursor
Understand the co-variance of x and y
Memory and fluency building (practice)
Distal precursor
Being able to locate two adjacent data points
Induction and refinement
Proximal precursor
Identify trend / rate of change
misconception: focus on only one variable
Interpolation & extrapolation
Understanding and sense making
Perform extrapolation or interpolations in a graph
misconception: believe that it is impossible to go beyond the range

FIGURE 5 | A specification diagram of the KSA-change model for one node/skill of interpolation/extrapolation in a graph in the HERA's KSA-model.

introduced in ECD is between a task model design based on a Proficiency model and a task-centered design (Mislevy et al., 1999). While in task-centered design, the primary emphasis is on creating the task with the target of inference defined only implicitly, as the tendency to do well on those tasks, in defining a task model based on a Proficiency (and Evidence) model, we make the connections and possible inferences *explicit from the start*, making the design easier to communicate, easier to modify, and better suited to principled generation of tasks (Mislevy et al., 1999, p. 23). Moreover, basing a task model on Proficiency and Evidence models allows us to consider reliability and validity aspects of task features, and particularly the cognitively or empirically based relevance of the task features. In other words, considerations of item reliability and validity guide the development of items to elicit the target observables and *only them* (minimizing added "noise"). This means that at the development stage of a task, all features of the task should stand to scrutiny regarding relevance to the latent KSA. As mentioned above, if reading ability is not relevant as part of the mathematics KSA, items or tasks that may impede students with lower reading skills should be avoided. Thus, defining a task model based on a Proficiency model resembles the relationship between the latent trait and its manifestation in observable behavior. The more the task relates to the target KSA, the better the inference from the observable to the latent KSA.

For assessment precision purposes per-se, there is no need to provide feedback to students; on the contrary, feedback can be viewed as interference in the process of assessment, and likewise scaffolds and hints introduce noise or interference to a single-point-in-time measurement. However, when the assessment tool is also intended for learning, the goal is to support learners when a weakness was identified, in order to help them gain the "missing" KSA. In the e-ECD we define a "Task-support model" that together with the original Task model creates the expanded-Task model (e-Task model). The Task-support model specifies the learning supports that are necessary and should be provided to learners in order to achieve KSA change. Similar to basing the Task model on the KSA model, the Task-support model is based on the KSA-change model. The supports may include customized feedback, hints and scaffolds, practice options, worked examples, explanations, or guidance to further tailored instruction derived from the latent learning processes specified in the KSA-change model. In other words, the supports are determined according to the focal knowledge *change*. We elaborate and illustrate on Task and Task-support models below.

The Assessment Layer Within the e-Task Model – The Task Model

The Task model provides a framework for describing *the situation* in which examinees are given the opportunity to exhibit their KSAs, and includes the specifications of the stimulus *materials*, *conditions* and *affordances*, as well as specifications for the *work product* (Mislevy et al., 1999, p. 19). The characteristics of the tasks are determined by the nature of the behaviors that provide evidence for the KSAs. Constructing a Task model from the latent KSA model involves considering the *cognitive aspect of task behavior*, including specifying the features of the situation, the internal representation of these features, and the connection between these representations and the problem-solving behavior the task targets. In this context, variables that affect task *difficulty* are essential to take into account. In addition, the Task model also includes features of task *management* and *presentation*.

Although the Task model is built off of the Proficiency model (or the KSA model in our notation), multiple Task models are possible in a given assessment, because each Task model may be employed to provide evidence in a different form, use different representational formats, or focus evidence on different aspects of proficiency. Similarly, the same Task model and work product can produce different evidence; i.e., different rules could be applied to the same work product, to allow inferences on different KSAs. Thus, it is necessary to define within each Task model the specific variables to be considered in the evidence rules (i.e., scoring rules; we elaborate on this in the next section).

Consider the abovementioned KSA from the HERA model: "*Perform an extrapolation using data from a graph*." As part of a scientific reasoning skills assessment, this skill is defined in a network of other skills related to understanding data representations, as seen in **Figure 5**. One possible Task model can be: "Given a graph with a defined range for the x-axis variable $[a,b]$ and y values corresponding to all x values in the range, find the *y*-value for an *x*-value outside the range." That is, we present the learner with a graph (defined by its *x*- and *y*- axes) and a function or paired coordinates (*x, y*) for a limited domain. The question then asks learners to predict the *y*-value of an *x* point which is outside the domain presented in the graph. Because extrapolation assumes the continuation of the trend based on the relationship between variables, a required characteristic of the question is to include this assumption, explicitly or implicitly *via* the context (e.g. stating other variables do not change, or the same experimental procedure was used for a new value). Articulating the assumption is part of the Task model. Another option for an extrapolation Task model could be: "Given a graph with two levels of the dependent variable, both showing a linear relationship with the x-variable (i.e., same relationship trend) but with different slopes, find the y-value for a third level of the dependent variable." That is, we present the learner with a graph with two linear relationships (two line-graphs), one for level *a* and one for level *b* (for example, *a, b* are levels of weight of different carts, and the linear relationship is between speed and time). The question then asks learners to predict the *y*-value for level *c* (*c* > *a, b;* larger weight car) for an *x*- point for which we know the *y*-values of level *a* and *b*; that is, extrapolation beyond the data presented. This Task model is more sophisticated than the first one, due to the complexity of the data representation, and thus is tapping into a higher level of the skill.

Another aspect is the operationalization of the Task model in a particular item. Given a Task model, the question can take the form of a direct non-contextualized (what we may also call a "naked") question, (e.g., asking about a value of *y* given a specific *x*), or it can be contextualized (or "wrapped") within the context and terminology of the graph (e.g., "suppose the researcher decided to examine the speed of a new cart that has greater weight, and suppose the trend of the results observed is maintained, what would you expect the new result to be?"). The "naked" and "dressed" versions of the question may involve change in the difficulty of the item; however, this change needs to be examined, to the extent that it is construct- relevant or irrelevant. If it is construct-relevant, then it should be included in the Task model as part of the specifications. Other factors may affect the difficulty as well – the type of graphic (bar-graph, line-graph, multiple lines, scatter plot) and the complexity of the relationships between variables (linear, quadratic, logarithmic, increasing, decreasing, one y-variable or more), the familiarity of the context of the task (whether this is a phenomenon in electricity, projectile motion, genetics, etc.), the complexity of the context (commonly understood, or fraught with misconceptions), the response options (multiple choice, or open-ended), the quality of the graph and its presentation (easy or hard to read, presented on a computer, smartphone or a paper, presented as a static graph or interactive where learners can plot points), etc. These factors and others need to be considered when specifying the Task model, and their relevance to the construct should be clearly articulated.

The Learning Layer Within the e-Task Model – The Task-Support Model

Tasks for assessment and tasks for learning differ in the availability of options that support learning. When we design tasks for learning, we need to consider the type of "help" or "teaching" that the task affords, with the same level of rigor that we put into the design of the task itself. The Task-support model thus specifies the learning supports that might be necessary and should be provided to students in order to achieve the desired KSA-change (i.e., increase in KSA). Similar to basing the task model on the KSA model, the Task-support model is based on the KSA-change model.

Making room for the specification of the task support *in connection* to the learning processes/goals (the focal KSA-change) is the innovative core of the proposed e-ECD and its significant contribution to the design of learning and assessment systems. Many learning systems include scaffolds or hints to accompany items and tasks, often determined by content experts or teacher experience and/or practices. These hints and scaffolds help answer the particular item they accompany, and may also provide "teaching," if transfer occurs to subsequent similar items. However, in the design process of the hints and scaffolds, often no explicit articulation is made regarding the intended effect of hints and scaffolds *beyond* the particular question, or in connection to the general learning goals. Often, the hints or scaffolds are task-specific; a breakdown of the task into smaller steps, thus decreasing the difficulty of the task. This is also reflected in the approach to assigning partial credit for an item that was answered correctly with hints, contributing less to the ability estimate (as evidence of lower ability; e.g., Wang et al., 2010). Specifying a Task-support model per each Task model dictates a *standardization* of the scaffolds and hints (and other supports) provided for a given task. How do we specify task supports connected to the focal KSA-change?

If for example, we define a particular (as part of the KSA-change model) learning model similar to the one depicted in **Figure 5**, we may provide as a task support a "pointer" to the precursors, in the form of a hint or a scaffold. Thus, the scaffolds are not a breakdown of the question to sub-steps, but rather each scaffold points to one of the precursor pieces of knowledge (initial, distal, or proximal precursor). In addition, since we defined the kind of knowledge change between each precursor, we can provide the corresponding support per each desired change. If the knowledge change is related to memory and fluency-building, we may provide more practice examples instead of the scaffold. Similarly, if the knowledge change is related to understanding and sense-making, we may provide an explanation or reasoning, or ask the student to provide the explanation or reasoning (self-explanation was found to be beneficial in some case, Koedinger et al., 2012). It may very well be the case that similar scaffolds will result from explicating a Task-support model following an e-ECD compared to not doing so, however in following this procedure, the design decisions are explicit and easy to communicate, justify, modify, replicate, and apply in a principled development of scaffolds.

Similarly, other features of task support, such as feedback, visuals, and links to a video or wiki page, can be supported by the articulation of the KSA-change and the connection between the two.

Let us illustrate specifying a Task-support model for the example item from HERA described in the previous section. Recall that the item targeted the latent KSA "*Perform an extrapolation using data from a graph*," and the task materials included a graph with a specified function, asking students to extrapolate a point beyond the given range (i.e., predict the value of *y* for a new *x-value*). Also, recall **Figure 5** that depicts the KSA-change model for this particular subskill. Given the proximal, distal, and initial precursors, we can now specify each scaffold to address each of these three precursor skills. Alternatively, we can decide to address only the closest precursor (the proximal) as a scaffold, and if that does not help with answering the question correctly, then refer the student to "learn" the more basic material (e.g., in a different section of the system, or by presenting items/content that target the initial and distal precursor skills). These decisions depend on the system design (e-Assembly model) and may vary from system to system.

As part of our development of the HERA system for scientific thinking skills, we developed an item model that can be used to collect evidence for both assessment and learning, termed an Assessment and Learning Personalized Interactive item (AL-PI). This item looks like a regular assessment item, and only after an incorrect response, the learners are given "learning options" to choose from. We offer three types of learning supports: (1) Rephrase – rewording of the question; (2) Break-it-down – providing the first step out of the multi-steps required to answer the question; and (3) Teach-me – providing a text and/or video explanation of the background of the question. **Figure 6** presents a screenshot of an AL-PI item from a task about height-restitution of a dropped-ball, targeting the skill of extrapolation.

Using the terminology above, the Rephrase-option provides the learner with another attempt at the question, with the potential of removing the construct irrelevance that may stem from the item-phrasing (for learners who did not understand what the question is asking them, due to difficulty with the wording). In this example, a Rephrase of the question is: "The question asks you to find the "Height attained" (the *y*-value) for a new *x*-value that does not appear on the graph" (see **Figure 6** upper panel). Note that the Rephrase is practically "undressing" (decontextualizing) the question, pointing out the "naked" form, or making the connection between the context and the decontextualized skill.

The second learning support is Break-it-down which takes the form of providing the first step to answer the question. In the example in **Figure 6** the Break-it-down states: "The first step to answer this question is to evaluate the rate of change in *y* as a function of a change in the *x*-variable" with additional marks and arrows on the graph to draw the leaner's attention where to look. The Break-it-down option may look like a hint, signaling to learners where to focus, and in our terminology, it refers to the proximal precursor (recall: proximal precursor = *identifying the rate of the change in the dependent variable as the independent variable changes*).

The third type of support that we offer in an AL-PI item is Teach-me. The Teach-me option in this case includes the following components: (1) a general statement about the skill; i.e., *a graph presents data for a limited number of values, yet we can estimate or predict about new values based on the trend in the data presented*; (2) an explanation of how to identify the trend in a graph, i.e., locating adjacent points; and (3) an illustration of how once the trend was identified, we can perform extrapolation.

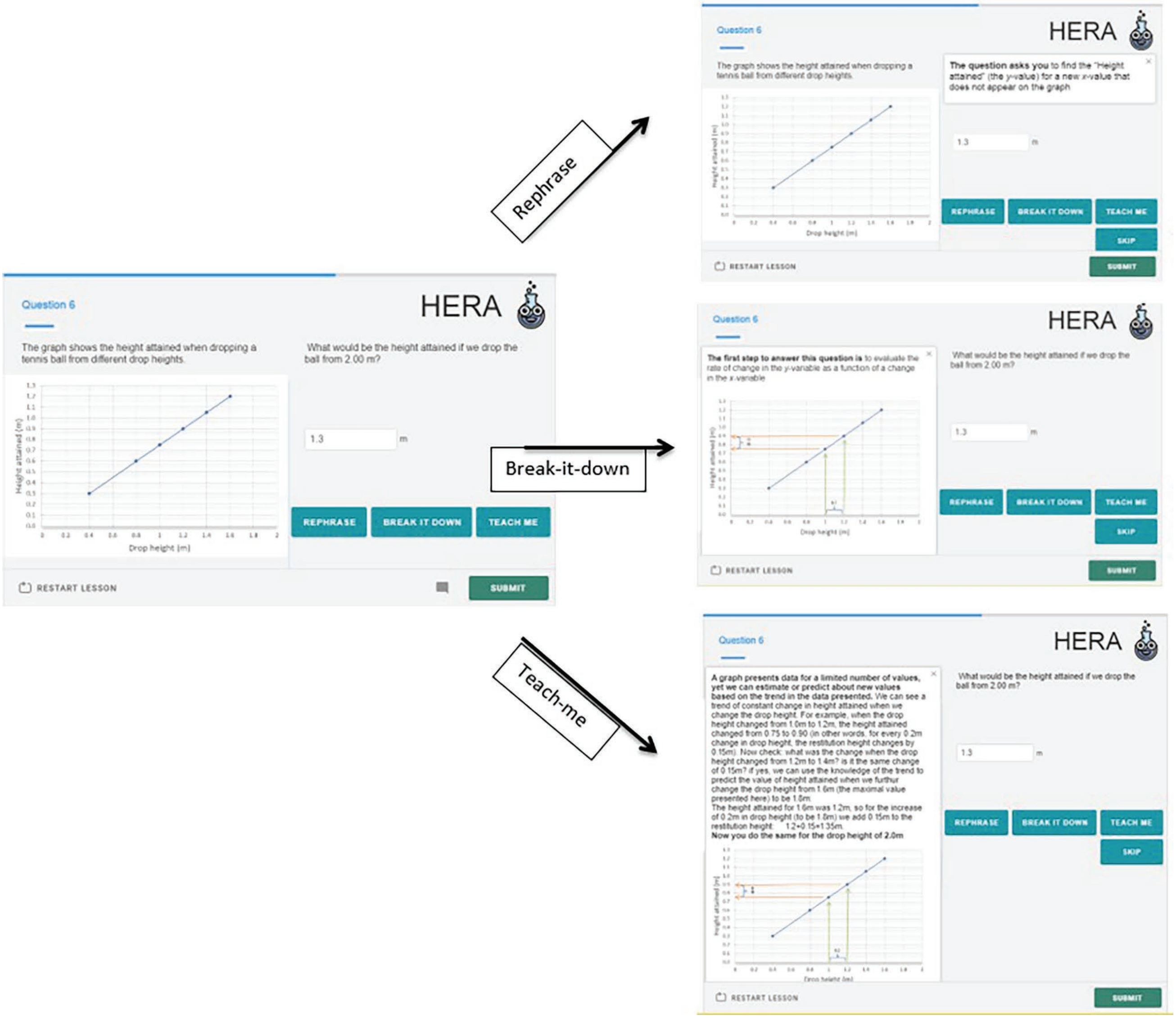
In our system we provide an illustration on a different value than the one in the question in order to avoid revealing the correct answer and leaving room for the learner to put mental effort into applying the method taught. In the Task-support model terminology and in relation to the KSA-change model, the Teach-me option addresses all three precursors.

Specifying the task support based on the learning goal and the desired change in KSA gives direction but does not limit the options. On the contrary, it enriches the space of the decision and opens-up new options. In addition, constructing task support by following the e-ECD framework gives rise to the hypothesis that this way of structuring scaffolds may enhance transfer, because the scaffolds do not address the particular question, but rather address the latent skill and its precursor skills. Empirical evidence of transfer is of course needed to examine this hypothesis.

Expanded Evidence Model

The links made between the e-Proficiency model and the e-Task model need explication of the statistical models that allow inferences from the work products on the tasks to the latent KSAs. In the ECD framework, the Evidence model specifies the links between the task's observables (e.g., student work product) and the latent KSAs targeted by that task (termed here as Observational-Evidence model). The Observational-Evidence model includes the *evidence rules* (scoring rubrics) and the *statistical models*. The Evidence model is the heart of the ECD, because it provides the "credible argument for how students' behaviors constitute evidence about targeted aspects of proficiency" (Mislevy et al., 1999, p. 2).

In a system designed for learning, data other than the work product is produced, i.e., the data produced out of the task support (e.g., hints and scaffolds usage), which may be called *process data*. The task support materials are created to foster learning; thus, learning systems should have a *credible argument* that these supports indeed promote learning. Partial evidence for that can be achieved by inferences about knowledge or what students know and can do from their work product in the system, *following* and as *a result of* the use of the supports, and this can be obtained by the statistical models within the Evidence model. However, the efficacy of the task supports themselves (i.e., which support helps the most in which case), and drawing inferences from scaffolds and hint usage about "learning behavior" or "learning processes" (as defined in the



The graph shows the height attained when dropping a tennis ball from different drop heights.

What would be the height attained if we drop the ball from 2.00 m?

The question asks you to find the "Height attained" (the *y*-value) for a new *x*-value that does not appear on the graph

The first step to answer this question is to evaluate the rate of change in the *y*-variable as a function of a change in the *x*-variable

A graph presents data for a limited number of values, yet we can estimate or predict about new values based on the trend in the data presented. We can see a trend of constant change in height attained when we change the drop height. For example, when the drop height changed from 1.0m to 1.2m, the height attained changed from 0.75 to 0.90 (in other words, for every 0.2m change in drop hiegth, the restitution height changes by 0.15m). Now check: what was the change when the drop height changed from 1.2m to 1.4m? is it the same change of 0.15m? if yes, we can use the knowledge of the trend to predict the value of height attained when we furthur change the drop height from 1.6m (the maximal value presented here) to be 1.8m.
The height attained for 1.6m was 1.2m, so for the increase of 0.2m in drop height (to be 1.8m) we add 0.15m to the restitution height: 1.2+0.15=1.35m.
Now you do the same for the drop height of 2.0m

FIGURE 6 | An example of an Assessment & Learning Personalized & Interactive item (AL-PI item) from the HERA system.

KSA-change model) may need new kind of models and evidence. The Transitional-Evidence model within the e-Evidence model addresses the data produced from the task support.

The Assessment Layer Within the Evidence Model – The Observational-Evidence Model

In the original ECD, the Observational-Evidence model addresses the question of how to operationalize the conceptual target competencies defined by the Proficiency model, which are essentially latent, in order to be able to validly infer from overt behaviors about those latent competencies. The Observational-Evidence model includes two parts. The first contains the scoring rules, which are ways to extract a "score" or an *observable variable* from student actions. In some cases, the scoring rule is simple, as in a multiple-choice item, in which a score of 1 or 0 is obtained corresponding to a correct or incorrect response. In other cases, the scoring rule might be more complex, as in performance assessment where student responses produce what we call "process data" (i.e., a log file of recorded actions on the task). A scoring rule for process data can take the form of grouping a sequence of actions into a "cluster" that may indicate a desired strategy, or a level on a learning progression that the test is targeting. In such an example, a scoring rule can be defined such that a score of 1 or 0 is assigned corresponding to the respective strategy employed, or the learning progression level achieved. Of course, scoring rules are not confined to dichotomous scores and they can also define scores between 0 and 1, continuous (particularly when the scoring rules relies on response time) or ordered categories of 1-to-*m*, for *m* categories (polytomous scores).

The second part of the Observational-Evidence model contains the statistical model. The statistical model expresses how the scores (as defined by the scoring rules) depend, probabilistically, on the latent competencies (the KSAs). This dependency is probabilistic, that is, the statistical model defines the probability of certain "scores" (observables) given specific latent competencies (combination of values on the KSAs). In other words, at the point in time at which the student is working within the system, that student is in a "latent state" of knowledge, and given that latent state, there is a certain probability for the observable variables, which if observed, are evidence for the latent ability. However, all we have are the student observable variables, and what we need is a psychometric model that allows us to do the *reverse inference* from the given observables to the latent competencies.

There are various statistical models that can be used here. Since we are talking about an assessment and learning system, let us consider a multi-dimensional latent competency, i.e., multiple skills are targeted by the system both for assessment and learning. If we assume the latent competencies to be continuous, we can use a multi-dimensional Item Response Theory models (e.g., MIRT; Reckase, 2009) or Bayes-net models (Pearl, 1988, 2014; Martin and VanLehn, 1995; Chang et al., 2006; Almond et al., 2015). In the case where the latent competencies are treated as categorical with several increasingly categories of proficiency in each (e.g., low-, medium-, and high-level proficiency, or mastery/non-mastery levels), we can use diagnostic classification models (DCM; Rupp et al., 2010b). What these models enable is to "describe" (or model) the relationship between the latent traits and the observables in a probabilistic way, such that the probability of a certain observable, given a certain latent trait, is defined and therefore allow us to make the *reverse* inference – to estimate the probability of a certain level of a latent trait given the observable.

In order to make the link between the items/tasks (the stimuli to collect observables) and the latent KSAs, we can use what is called a Q-matrix (Tatsuoka, 1983). A Q-matrix is a matrix of <items × skills> (items in the rows; skills in the columns), defining for each item which skills it is targeting. The Q-matrix plays a role in the particular psychometric model, to determine the probability of answering an item correctly given the combination of skills (and whether all skills are needed, or some skill can compensate for others; non-compensatory or compensatory model, respectively). The Q-matrix is usually determined by content experts, but it can also be learned from the data (e.g., Liu et al., 2012).

Recent developments in the field of psychometrics have expanded the modeling approach to also include models that are data driven, but informed by theory, and is referred to as Computational Psychometrics (von Davier, 2017). Computational Psychometrics is a framework that includes complex models such as MIRT, Bayes-net and DCM, which allow us to make inferences about latent competencies; however, these models may not define *a priori* the scoring rules, but rather allow for a combination of the expert-based scoring rules with those that are learned from the data. In particular, the supervised algorithms – methodologies used in machine learning (ML) – can be useful for identifying patterns in the complex logfile data. These algorithms classify the patterns by skills using a training data set that contained the correct or theory-based classification. The word supervised here means that the "correct responses" were defined by subject-matter experts and that the classification algorithm learns from these data that were correctly classified to extrapolate to new data points.

In a learning and assessment system, the Observational-Evidence model may also take into account the scaffolds and hints usage to infer about the KSA model. Since the scaffolds and hints reduce the difficulty of the items/tasks, they also change their evidentiary value of the observables. This can be done *via* either using only responses without hint usage to model KSA or applying a partial credit scoring rule for items that were answered correctly with hints, thus assigning them less credit as a reflection of their evidentiary value (e.g., Wang et al., 2010; Bolsinova et al., 2019a,b).

To summarize, any and all statistical models that allow us to define the connection between overt observables and latent competencies can be used in the Observational-Evidence model.

The Learning Layer Within the Evidence Model – The Transitional-Evidence Model

Similar to the way the Observational-Evidence model connects the Task model back to the KSA model, the Transitional-Evidence model uses the task supports data to infer about learning, and to link back to the KSA-change model. Recall that the KSA-change model includes pedagogical principles which are reflected in the task supports. Similar to the assessment layer of the Evidence model, the Transitional-Evidence model also includes two parts: the scoring rules and the statistical models.

The scoring rules define the *observable variables* of the Transitional-Evidence model. If task supports are available by choice, student choice behavior can be modeled to make inferences about their learning strategies. The data from the task supports usage (hints, scaffolds, videos, simulations, animations, etc.) as well as number of attempts or response time, should first be coded (according to a scoring or evidence rule) to define which of them should count and in what way. As before, scoring rules can be defined by human experts or can be learned from the data.

The statistical models in the Transitional-Evidence model need to be selected, such that they allow us to infer about *change* based on observables over time. A popular stochastic model for characterizing a changing system is a Markov model (cf. Norris, 1998). In a Markov model, transition to the next state depends only on the current state. Because the focus here is on latent competencies, the appropriate model is then a hidden Markov model (HMM; e.g., Visser et al., 2002; Visser, 2011), and specifically an input-output HMM (Bengio and Frasconi, 1995). A HMM would allow us to infer about the efficacy of the learning supports in making a change in the

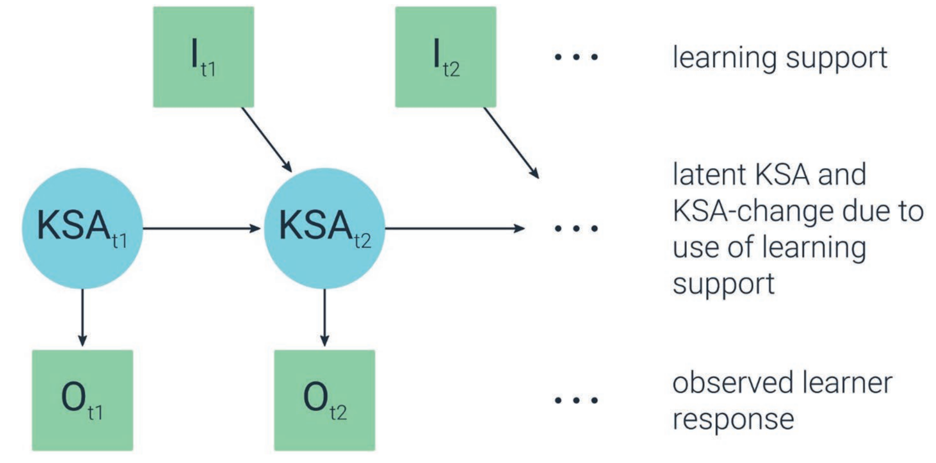

I_{t1}
I_{t2}
learning support
KSA_{t1}
KSA_{t2}
latent KSA and KSA-change due to use of learning support
O_{t1}
O_{t2}
observed learner response

FIGURE 7 | An input-output hidden Markov model (HMM).

latent state (proficiency level). In addition, the input-output HMM will allow us to make the association between learning materials (as input) and the change in KSA (latent) based on the observables (output), to estimate the contribution (efficacy) of each particular support to the desired change in proficiency (i.e., learning). **Figure 7** illustrates this model for a single latent skill (KSA at time t1 and t2), a single observation (O at time t1 and t2) and a single learning support (l at time t1 and t2). The observation dependency on the skill (i.e., O given KSA; the arrow/link from KSA to O) is modeled by the Observational-Evidence model (the model from the original ECD), while the skill dependency on the learning support (i.e., KSA given l; the arrow/link from l to KSA) is modeled by the Transitional-Evidence model.

Working with the above example, let us assume a student does not know how to identify a data trend from a graph, and thus cannot extrapolate a new data point (incorrectly answers a question that requires extrapolation). Suppose a task support is provided, such that it draws the student's attention to the pattern and trend in the data. We now want to estimate the contribution of this support in helping the student learn (and compare this contribution to other task supports). We have the following observables: the student's incorrect answer in the first attempt, the student's use of the particular task support, and the student's revised answer in the second attempt (whether correct or not). Using an input-output HMM will allow us to estimate the probability of transitioning from the incorrect to the correct latent state (or in other cases from low proficiency to high proficiency), given the use of the task support. Of course, the model will be applied across questions and students in order to infer about latent state.

The above example of a single latent skill can be extended to a map of interconnected skills using dynamic Bayesian network (DBN; Murphy and Russell, 2002). DBN generalizes HMM by allowing the state space to be represented in a factored form instead of as a single discrete variable. DBN extends Bayesian networks (BN) to deal with changing situations.

How do we link the learning materials (defined in the Task-support model) to the learning processes/goals (defined in the KSA-change model)? Similar to the Q-matrix in the Observational-Evidence model, here too we need a matrix that links the learning materials (task supports) with the associated skills-change. We can use an S-matrix (Chen et al., 2018), which is a matrix of <supports × skills> (supports in the rows; skills in the columns), defining for each support which skills/process it can improve. In that sense, and similar to the Q-matrix, an S-matrix is a collection of "evidence" that explicate the connection between the supports and the desired learning shifts. For example, *providing a worked example* is a learning support that may be connected to several knowledge shifts (corresponding to subskills in the learning models), and providing opportunities for practice is another learning support that may be connected to different desired knowledge shifts (corresponding to different subskills). The S-matrix will specify these connections. The S-matrix will then play a role in the HMM, to determine the probability that a particular knowledge shift (learning process) occurred given the particular learning supports. Similar to the Q-matrix, the S-matrix should be determined by content experts, and/or learned or updated from the data.

THE e-ASSEMBLY MODEL

In the original ECD, the Assembly model determines how to put it all together and specifies the conditions needed for obtaining the desired reliability and validity for the assessment. In other words, it determines the *structure* of the test, the number and the mix of the desired items/tasks. The Assembly model is directly derived from the Proficiency model, such that it ensures, for example, the appropriate representation of all skills in the map. Going back to the HERA example and the KSA-model in **Figure 3**, if we were to build an assessment with those target skills, we would have to ensure that we sample items/tasks for each of the skills and subskills specified on the map, and the Assembly model will specify how much of each.

For the expanded ECD, we do not create a parallel model to the Assembly model as we did for the three core models, because in a blended learning and assessment system we do not assemble the assessment separately and the learning separately. Rather, in the process of developing a system, after we specified the six core models of the e-ECD, we assemble it all together in what we call the e-Assembly model.

The role of e-Assembly model is to specify how to put it all together. It will include the specifications of number and mix of items/tasks, but it will also include how and when to present the learning support materials. This can be seen as determining how to switch between the "assessment" mode of the system and the "learning" mode of the system.

The e-Assembly model provides an opportunity to take into account additional pedagogical principles that are relevant to the *combination* of items and tasks, such as the importance of reducing cognitive load for learning; focusing on one skill at a time; gradual increased difficulty presentation; adaptive presentation of content, among others. Conditions to ensure the validity of the system may also specify pedagogical principles such as learning *via* real-world authentic tasks or learning by doing, as well as learner engagement factors, as relevant. Pedagogical Content Knowledge principles that include knowledge of student *misconceptions* regarding specific phenomena, if articulated as part of the KSA and KSA-change model, should be also considered here in selecting and designing tasks, such that the misconceptions are either accounted for or avoided so the KSAs can be validly addressed.

The e-Assembly model is also the place to take into account considerations from other relevant approaches, such as the learner-centered design approach (LCD; Soloway et al., 1994; Quintana et al., 2000), which argue that student engagement and constructivist theories of learning should be at the core of a *computerized* learning system. Adopting such an approach will affect the combination and/or navigation through the system. For example, the system may guide students to be more *active* in trying out options and *making choices* regarding their navigation in the system.

An important aspect of systems for learning and assessment is whether they are adaptive to student performance and in what way. This aspect within the e-Assembly model ties directly to the e-Evidence model. The statistical models in the Evidence model are also good candidates for determining the adaptive algorithm in adaptive assessments. For example, if a 2PL IRT model is used to estimate ability; this model can also be used to select the items in a Computer Adaptive Test (CAT), as is often done in large-scale standardized tests that are adaptive (e.g., the old version of the GRE). Similarly, if a Bayes-net is used to estimate the map of KSAs, then the selection of items or tasks can be done based on the Bayes-net estimates of skills. Similarly, we can use the DCM to identify weakness in a particular skill and thus determine the next item that targets that particular weakness. This is true for any other model, also including data-driven models, because the purpose of the models is to provide a valid way to estimate KSAs, and once this is done, adaptivity within the system can be determined accordingly.

The learning aspect of the system is motivated by the goal to maximize learners' gain and thus needs a more comprehensive adaptivity, or what is often called "recommendation model." A recommendation model does not only determine the next item to be presented but it also determines which instructional or training material to recommend or present to the learner. A good recommendation model makes full use of all available information about both the learner and the instructional materials to maximize the KSA gain for the learner. If we have a way to estimate (measure) the gain for the learner, we can feed this information to the recommendation engine to determine the adaptivity in the form of the next task support and/or training and instructional material needed. Thus, the additional layer of an evidence model for the learning materials (i.e., the statistical models for estimating the efficacy of the task supports) provides a good candidate model for the recommendation engine. Which materials were already used by the learner (which ones were chosen/preferred), which supports are found more effective for that particular learner, which skill is currently in focus and which supports are most effective for that particular skill (e.g., practice, explained example, video lecture, simulation demonstration, providing instructional material for a prior/prerequisite skill, etc.) are some of the decisions needed to be made by a recommendation engine, and these decisions rely on the statistical models that were used to evaluate and provide evidence for the efficacy of the task support and instructional materials.

CONCLUSION AND FUTURE STEPS

In this paper, we propose a new way to fuse learning and assessment at the design stage. Specifically, we propose an expanded framework we developed to aid with the creation of a system for blended assessment and learning. We chose the ECD framework as a starting point because this is a comprehensive and rigorous framework for the development of assessments and underlies the development of tests for most testing organizations. Incorporating learning aspects, both learning goals and learning processes, in the ECD framework is challenging, because of fundamental differences in the assumptions and approaches of learning and assessment. Nevertheless, we showed that the unique structure of Proficiency, Task, and Evidence models lends itself to creating parallel models for consideration of the corresponding aspect of learning within each model.

We are currently applying this framework in our work. In future work, we hope to show examples of the learning and assessment system that we build following the e-ECD framework. We are also working to incorporate other elements into the framework, primarily the consideration of motivation, meta-cognition, and other non-cognitive skills. Since learners' engagement is a crucial element in a learning system, we can think of a way to incorporate elements that enhance engagement as part of the assembly of the system, by using reward system or gamification in the form of points, coins, badges, etc. Adding gamification or engagement-enhancing elements into a system does not currently have a designated model within the e-ECD. We are working to find a way to incorporate these elements into the framework.

AUTHOR CONTRIBUTIONS

MA-A and AAvD contributed to the conception of the framework. MA-A contributed to the conception and specifications of the new models, and AAvD contributed to the CP component. SW and JT contributed to the e-Task model. BD contributed to the e-Evidence model. The authors would like to thank the reviewers for substantial contribution.

FUNDING

This work has been done as part of a research initiative at ACTNext, by ACT, Inc. No external funds or grants supported this study.

REFERENCES

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., and Williamson, D. M. (2015). *Bayesian networks in educational assessment*. (New York, NY: Springer).

Anderson, J. R., Corbett, A. T., Koedinger, K. R., and Pelletier, R. (1995). Cognitive tutors: lessons learned. *J. Learn. Sci.* 4, 167–207. doi: 10.1207/s15327809jls0402_2

Arieli-Attali, M., and Cayton-Hodges, G. (2014). Expanding the CBAL™ mathematics assessments to elementary grades: the development of a competency model and a rational number learning progression. *ETS Res. Rep. Ser.* 2014, 1–41. doi: 10.1002/ets2.12008

Arieli-Attali, M., Wylie, E. C., and Bauer, M. I. (2012). "The use of three learning progressions in supporting formative assessment in middle school mathematics" in *Annual meeting of the American Educational Research Association.* (Vancouver, Canada).

Attali, Y., and Arieli-Attali, M. (2014). Gamification in assessment: do points affect test performance? *Comp. Educ.* 83, 57–63. doi: 10.1016/j.compedu.2014.12.012

Bengio, Y., and Frasconi, P. (1995). "An input output HMM architecture" in *Advances in Neural Information Processing Systems.* eds. M. I. Jordan, Y. LeCun, and S. A. Solla (Cambridge, MA, USA: MIT Press), 427–434.

Bolsinova, M., Deonovic, B., Arieli-Attali, M., Settles, B., Hagiwara, M., Von Davier, A., et al. (2019a). Hints in adaptive learning systems: consequences for measurement. Paper presented at the annual meeting of the National Council of Measurement in Education (NCME). Toronto, Canada.

Bolsinova, M., Deonovic, B., Arieli-Attali, M., Settles, B., Hagiwara, M., Von Davier, A., et al. (2019b under review). Measurement of ability in adaptive learning and assessment systems when learners use on-demand hints. *Educ. Psychol. Meas*.

Chang, K. M., Beck, J., Mostow, J., and Corbett, A. (2006). "A Bayes net toolkit for student modeling in intelligent tutoring systems" in *International Conference on Intelligent Tutoring Systems.* (Berlin, Heidelberg: Springer), 104–113.

Chen, Y., Li, X., Liu, J., and Ying, Z. (2018). Recommendation system for adaptive learning. *Appl. Psychol. Meas.* 42, 24–41. doi: 10.1177/0146621617697959

Conrad, S., Clarke-Midura, J., and Klopfer, E. (2014). A framework for structuring learning assessment in an educational massively multiplayer online educational game – experiment centered design. *Int. J. Game Based Learn.* 4, 37–59. doi: 10.4018/IJGBL.2014010103

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: application to abstract reasoning. *Psychol. Methods* 3, 300–396.

Feng, M., Hansen, E. G., and Zapata-Rivera, D. (2009a). "Using evidence centered design for learning (ECDL) to examine the ASSISTments system" in *Paper presented in the annual meeting of the American Educational Research Association* (*AERA*). (San Diego, California).

Feng, M., Heffernan, N. T., and Koedinger, K. R. (2009b). Addressing the assessment challenge in an intelligent tutoring system that tutors as it assesses. *J. User Model. User Adapt Interact.* 19, 243–266. doi: 10.1007/s11257-009-9063-7

Furtak, E. M., Thompson, J., Braaten, M., and Windschitl, M. (2012). "Learning progressions to support ambitious teaching practices" in *Learning progressions in science: Current challenges and future directions.* eds. A. C. Alonzo, and A. W. Gotwals (Rotterdam: Sense Publishers), 405–433.

Grover, S., Bienkowski, M., Basu, S., Eagle, M., Diana, N., and Stamper, J. (2017). "A framework for hypothesis-driven approaches to support data-driven learning analytics in measuring computational thinking in block-based programming" in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (ACM). Vancouver, BC, Canada, 530–531.

Heffernan, N., and Heffernan, C. (2014). The ASSISTments ecosystem: building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *Int. J. Artif. Intell. Educ.* 24, 470–497. doi: 10.1007/s40593-014-0024-x

Kim, Y. J., Almond, R. G., and Shute, V. J. (2016). Applying evidence-centered design for the development of game-based assessments in physics playground. *Int. J. Test.* 16, 142–163. doi: 10.1080/15305058.2015.1108322

Kingston, N. M., Karvonen, M., Thompson, J. R., Wehmeyer, M. L., and Shogren, K. A. (2017). Fostering inclusion of students with significant cognitive disabilities by using learning map models and map-based assessments. *Inclusion* 5, 110–120. doi: 10.1352/2326-6988-5.2.110

Koedinger, K. R., Corbett, A. T., and Perfetti, C. (2012). The knowledge-learning-instruction framework: bridging the science-practice chasm to enhance robust student learning. *Cogn. Sci.* 36, 757–798. doi: 10.1111/j.1551-6709.2012.01245.x

Koehler, M., and Mishra, P. (2009). What is Technological Pedagogical Content Knowledge (TPACK)? *Contemp. Issues Technol. Teach. Educ.* 9, 60–70. doi: 10.1177/002205741319300303

Liu, J., Xu, G., and Ying, Z. (2012). Data-driven learning of Q-matrix. *Appl. Psychol. Meas.* 36, 548–564. doi: 10.1177/0146621612456591

Luecht, R. M. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *J. Appl. Test. Technol.* 14, 1–38. Retrieved from: http://jattjournal.com/index.php/atp/article/view/45254

Martin, J., and VanLehn, K. (1995). Student assessment using Bayesian nets. *Int. J. Hum. Comput. Stud.* 42, 575–591. doi: 10.1006/ijhc.1995.1025

Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. *Mil. Med.* (*special issue on simulation, H. O'Neil, Ed.*) 178, 107–114. doi: 10.7205/MILMED-D-13-00213

Mislevy, R. J., Almond, R. G., and Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Res. Rep. Ser.* 2003. Princeton, NJ. doi: 10.1002/j.2333-8504.2003.tb01908.x

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., and Levy, R. (2012). Design and discovery in educational assessment: evidence-centered design, psychometrics, and educational data mining. *JEDM J. Educ. Data Min.* 4, 11–48. Retrieved from: https://jedm.educationaldatamining.org/index.php/JEDM/article/view/22

Mislevy, R. J., Steinberg, L. S., and Almond, R. G. (1999). "On the roles of task model variables in assessment design." in Paper presented at the Conference "Generating Items for Cognitive Tests: Theory and Practice" (Princeton, NJ).

Mislevy, R. J., Steinberg, L. S., Almond, R. G., and Lukas, J. F. (2006). "Concepts, terminology, and basic models of evidence-centered design" in *Automated scoring of complex tasks in computer-based testing*. eds. D. M. Williamson, R. J. Mislevy, and I. I. Bejar (New York, NY), 15–47.

Murphy, K. P., and Russell, S. (2002). Dynamic Bayesian networks: Representation, inference and learning. Doctoral dissertation. Berkeley: University of California. Available at: https://www.cs.ubc.ca/~murphyk/Thesis/thesis.html (Accessed November 4, 2019).

National Research Council (NRC) (2007). *Taking science to school: Learning and teaching science in grades K-8*. (Washington, DC: The National Academies Press).

Nichols, P., Ferrara, S., and Lai, E. (2015). “Principled design for efficacy: design and development for the next generation tests” in *The next generation of testing: Common core standards, smarter-balanced, PARCC, and the nationwide testing movement*. ed. R. W. Lissitz (Charlotte, NC: Information Age Publishing), 228–245.

Nichols, P., Kobrin, J. L., Lai, E., and Koepfler, J. (2016). “The role of theories of learning and cognition in assessment design and development” in *The handbook of cognition and assessment: Frameworks, methodologies, and applications.* 1st edn. eds. A. A. Rupp and J. P. Leighton (Massachusetts, USA: John Wiley & Sons, Inc.), 15–40.

Norris, J. R. (1998). *Markov chains.* (New York, NY: Cambridge University Press).

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. (San Francisco, CA: Morgan Kaufmann Publishers, Inc.).

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. (Elsevier).

Pelánek, R. (2017). Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Model. User Adap. Inter.* 27, 313–350. doi: 10.1007/s11257-017-9193-2

Posner, G. J., Strike, K. A., Hewson, P. W., and Gertzog, W. A. (1982). Accommodation of a scientific conception: toward a theory of conceptual change. *Sci. Educ.* 66, 211–227. doi: 10.1002/sce.3730660207

Quintana, C., Krajcik, J., and Soloway, E. (2000). “Exploring a structured definition for learner-centered design” in *Fourth International Conference of the Learning Sciences*. eds. B. Fishman and S. O’Connor-Divelbiss (Mahwah, NJ: Erlbaum), 256–263.

Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N. T., Koedinger, K. R., Junker, B., et al. (2005). “The Assistment project: blending assessment and assisting” in *Proceedings of the 12th Artificial Intelligence in Education*. eds. C. K. Looi, G. McCalla, B. Bredeweg and J. Breuker (Amsterdam: ISO Press), 555–562.

Reckase, M. D. (2009). “Multidimensional item response theory models” in *Multidimensional item response theory.* ed. M. D. Reckase (New York, NY: Springer), 79–112.

Rupp, A. A., Gushta, M., Mislevy, R. J., and Shaffer, D. W. (2010a). Evidence-centered design of epistemic games: measurement principles for complex learning environments. *J. Technol. Learn. Assess.* 8. Retrieved from: https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1623

Rupp, A. A., Templin, J. L., and Henson, R. A. (2010b). *Diagnostic measurement: Theory, methods, and applications*. (New York, NY: Guilford Press).

Shute, V. J., Hansen, E. G., and Almond, R. G. (2008). You can’t fatten A hog by weighing It–Or can you? evaluating an assessment for learning system called ACED. *Int. J. Artif. Intell. Edu*. 18, 289–316. https://content.iospress.com/articles/international-journal-of-artificial-intelligence-in-education/jai18-4-02

Soloway, E., Guzdial, M., and Hay, K. (1994). Learner-centered design: the challenge for HCI in the 21st century. *Interactions* 1, 36–48. doi: 10.1145/174809.174813

Straatemeier, M. (2014). Math garden: a new educational and scientific instrument. *Education* 57, 1813–1824. PhD thesis. ISBN9789462591257.

Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory. *J. Educ. Meas.* 20, 345–354. doi: 10.1111/j.1745-3984.1983.tb00212.x

Ventura, M., and Shute, V. (2013). The validity of a game-based assessment of persistence. *Comput. Hum. Behav.* 29, 2568–2572. doi: 10.1016/j.chb.2013.06.033

Visser, I. (2011). Seven things to remember about hidden Markov models: a tutorial on Markovian models for time series. *J. Math. Psychol.* 55, 403–415. doi: 10.1016/j.jmp.2011.08.002

Visser, I., Raijmakers, M. E., and Molenaar, P. (2002). Fitting hidden Markov models to psychological data. *Sci. Program.* 10, 185–199. doi: 10.1155/2002/874560

von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. *J. Educ. Meas.* 54, 3–11. doi: 10.1111/jedm.12129

Wang, Y., Heffernan, N. T., and Beck, J. E. (2010). “Representing student performance with partial credit” in *Proceeding of a conference: 3rd International Conference on Educational Data Mining*. eds. R. S. J. d. Baker, A. Merceron, and P. I. Pavlik Jr. (Pittsburgh, PA, USA), 335–336.

Wilson, M. (2009). Measuring progressions: assessment structures underlying a learning progression. *J. Res. Sci. Teach.* 46, 716–730. doi: 10.1002/tea.20318

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 07 May 2019
doi: 10.3389/fpsyg.2019.01010

Development of a Computerized Adaptive Testing for Internet Addiction

Yong Zhang, Daxun Wang, Xuliang Gao*, Yan Cai* and Dongbo Tu**

School of Psychology, Jiangxi Normal University, Nanchang, China

Internet addiction disorder has become one of the most popular forms of addiction in psychological and behavioral areas, and measuring it is growing increasingly important in practice. This study aimed to develop a computerized adaptive testing to measure and assess internet addiction (CAT-IA) efficiently. Four standardized scales were used to build the original item bank. A total of 59 polytomously scored items were finally chosen after excluding 42 items for failing the psychometric evaluation. For the final 59-item bank of CAT-IA, two simulation studies were conducted to investigate the psychometric properties, efficiency, reliability, concurrent validity, and predictive validity of CAT-IA under different stopping rules. The results showed that (1) the final 59 items met IRT assumptions, had high discrimination, showed good item-model fit, and were without DIF; and (2) the CAT-IA not only had high measurement accuracy in psychometric properties but also sufficient efficiency, reliability, concurrent validity, and predictive validity. The impact and limitations of CAT-IA were discussed, and several suggestions for future research were provided.

Keywords: internet addiction, computer adaptive testing, item response theory, questionnaire, CAT-IA

Edited by:
Samuel Greiff,
University of Luxembourg,
Luxembourg

Reviewed by:
Ioannis Tsaousis,
University of Crete, Greece
Roger Ho,
National University of Singapore,
Singapore

***Correspondence:**
Daxun Wang
447951689@qq.com
Xuliang Gao
gaoxuliang8817@qq.com
Yan Cai
cy1979123@aliyun.com
Dongbo Tu
tudongbo@aliyun.com

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: *30 October 2018*
Accepted: *16 April 2019*
Published: *07 May 2019*

Citation:
Zhang Y, Wang D, Gao X, Cai Y and Tu D (2019) Development of a Computerized Adaptive Testing for Internet Addiction. Front. Psychol. 10:1010. doi: 10.3389/fpsyg.2019.01010

INTRODUCTION

Internet addiction (IA) disorder is now recognized as one of the most popular forms of addiction in psychological and behavioral areas. According to a report released by the International Telecommunication Union (2016), with the rapid development of advanced mobile networks, the number of users over the last 3 years has climbed to nearly four billion people, which is equivalent to 47% of the global population. Although the internet brings many benefits, excessive access to the network can lead to internet addiction (IA). A recent meta-analysis reported that the global prevalence of IA is 30.1% among university students pursuing a professional degree (Zhang et al., 2018). In Asia, the prevalence of IA ranged from 6.2% in Japanese adolescents to 21% in Filipino adolescents (Mak et al., 2014b). IA is associated with sleep disturbance (Zhang et al., 2017), poor quality of life (Tran et al., 2017a), and other psychiatric illnesses (Ho et al., 2014). Therefore, the assessment and prevention of IA are particularly important in practice. IA symptoms have been evaluated primarily by questionnaires that have been developed based on classical test theory. The commonly used questionnaires include the Internet Addiction Test (IAT; Young, 1998), Generalized Problematic Internet Use Scale (GPIUS; Caplan, 2002), Gaming Addiction Scale (GAS; Lemmens et al., 2009), and Revised Chen Internet Addiction Scale (CIAS-R; Mak et al., 2014a). The current questionnaires classify IA symptoms into loss of control or of time management (Tran et al., 2017b), craving and social problems (Lai et al., 2013). Although these questionnaires are frequently used in practice, they have certain weaknesses. One of the most notable drawbacks is that participants must finish all of the questionnaire items. However, many items may be "off target" for different test takers (Fliege et al., 2005). For participants with high ability levels, easy items have less contribution to measuring their actual ability level, and as such, these items may be redundant or unnecessary. Meanwhile, for participants with low ability levels, the requirement of responding to the difficult items results in the difficulty to measure their actual ability level. Therefore, it is essential to have a more effective method to evaluate IA.

One way to deal with the above issues is through computerized adaptive testing (CAT), which is a new kind of test that uses item response theory (IRT) to establish an item bank, and then automatically selects items according to the current theta of each participant, and finally estimates the ability of each test taker (Almond and Mislevy, 1999). In CAT, the test-taker continues to take test items until his/her estimated θ reaches a predefined level of precision, as indicated by its standard error. Compared with a linear test, CAT cannot only present items, input answers, and automatically score through the computer but also automatically select the most appropriate items for each responder according to the different answers to items, and then finally reach the most appropriate estimation of ability.

Many studies have shown that a CAT program has several advantages over paper-and-pencil questionnaires. Flens et al. (2016) revealed that compared with paper-and-pencil questionnaires, the number of used items based on CAT procedures decreases by 26–44%. Linacre (2000) pointed out that CAT programs can improve validation, reduce individuals' burden, and have more excellent measurement precision. In addition, with the selection of items based on a respondent's current theta, the floor and ceiling effects can be decreased in CAT procedures (Revicki and Cella, 1997). Further, the development of CAT procedures improves clinical assessment. However, CAT also has a number of disadvantages: high costs of research and development, complex technical requirements, and the need for timely maintenance of the item bank to prevent items from leaking in advance (Tan et al., 2018). Nonetheless, the virtues of a CAT program importantly overweigh the defects.

Initially, the development and applications of CAT programs mainly occurred in intelligence and ability testing (e.g., Tinsley, 1972; Ireland, 1977; Young, 1990). In recent years, many researchers have paid attention to the field of mental health. For example, Flens et al. (2017) used the IRT model to assess the Dutch-Flemish version of depression. Smits et al. (2011) established and evaluated CAT procedures for depression based on the Epidemiologic Studies-Depression scale. Walter et al. (2007) developed a German version of Anxiety CAT within IRT. However, to the best of our knowledge, the use of CAT to IA, a common disorder, has not been applied.

This study aimed to develop CAT to assess IA (CAT-IA) without loss of measurement precision. More specifically, this work addressed the following. First, a calibrated item bank with high psychometric qualities was developed. Second, in different stopping rules, we evaluated the psychometric properties, efficiency, reliability, and validity of CAT-IA via two CAT simulation studies. Third, we sought to extend the applications of CAT in the field of mental health and introduce IRT and CAT to readers who want to understand and apply adaptive testing.

MATERIALS AND METHODS

Participants

The total sample consisted of 1,368 participants. All of the participants were surveyed at different schools in China from June to September 2017. **Table 1** reveals the characteristics of the participants. The sample included 687 females (50.2%) and 681 men (49.8%). Their average age was 18.72 years (*SD* = 2.19, ranged from 12 to 28 years). The participants came from two regions: rural (58.9%) and urban (41.1%).

This study was conducted at the Research Center of Mental Health, Jiangxi Normal University, following the recommendations of psychometrics studies on mental health. It was approved by the Research Center of Mental Health, Jiangxi Normal University and the Ethics Committee of the Department of Psychology at Jiangxi Normal University. Written informed consent was obtained from all of the participants in accordance with the Declaration of Helsinki. Parental consent was also obtained for all participants under the age of 16 years.

Measures and the Initial Item Pool

The initial item pool of CAT-IA consisted of 101 items (see **Table 2**). These items were selected from four standardized scales: IAT (Young, 1998), GPIUS (Caplan, 2002), GAS (Lemmens et al., 2009), and Chinese Internet Addiction Test (CIAT; Huang et al., 2007). All of them used five-point Likert-type item scores (never, rarely, sometimes, often, always; scored with 1, 2, 3, 4, and 5, respectively). A higher cumulative sum in all of the items represented more severe symptoms of IA. Based on previous studies, 101 items from the four selected standardized scales could be classified into seven domains (Young, 1998; Caplan, 2002; Huang et al., 2007; Lemmens et al., 2009): salience, tolerance, mood modification, relapse, withdrawal, negative outcomes, and benefits (i.e., compared with offline, individuals are more likely to participate in social behavior online and surfing the internet can reduce negative emotions).

TABLE 1 | The characteristics of participants (*n* = 1,368).

Characteristics	% or years
Gender	
Female	50.2
Male	49.8
Age	
Mean	18.72
SD	2.19
Range	12–28
Region	
Rural	58.9
Urban	41.1

TABLE 2 | Items from four scales.

Scale	Number of items	Items
IAT	20	IAT-1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, and IAT-20
GPIUS	29	GPIUS-21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, and GPIUS-49
GAS	21	GAS-50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, and GAS-70
CIAT	31	CIAT-71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, and CIAT-101

IAT, Internet Addiction Test; GPIUS, Generalized Problematic Internet Use Scale; GAS, Gaming Addiction Scale; CIAT, Chinese Internet Addiction Test.

Item Bank Construction of CAT-IA

To obtain a high-quality item bank, psychometric evaluations were performed on the individuals' actual data as follows.

Step 1: Test the unidimensional assumption of the item pool.

Unidimensionality means that the test measures only one main latent trait; that is, responses on each item are affected by one main latent trait of the participants (Embretson and Reise, 2013). Both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were used to assess the unidimensional assumption. In EFA, the unidimensional assumption is deemed sufficient when the first factor explains at least 20% of the variance (Reckase, 1979), and the ratio of the explained variance in the first and second factor is greater than 4 (Reeve et al., 2007). The CFA of a single-factor was used to assess the unidimensional assumption. We adopted two indicators: factor loading and root mean square error of approximation (RMSEA) estimated by the weighted least square means and variance adjusted method using Mplus7.0 (Muthén and Muthén, 2012). According to the rule of thumb of Browne and Cudeck (1993), the model has a close fit, is fair or acceptable, mediocre, or poor if the RMSEA value is below 0.05, between 0.06 and 0.08, between 0.09 and 0.10, or above 0.10, respectively. We excluded items with factor loadings smaller than 0.4 because factor loadings below 0.4 could easily be over-interpreted (Nunnally, 1978).

Step 2: Select the appropriate IRT model according to the test-level model-fit indices.

Selecting the appropriate model is one of the most important procedures to make valid inferences. In this study, four commonly used polytomous IRT models were considered: Graded Response Model (GRM; Samejima, 1969), Generalized Partial Credit Model (GPCM; Muraki, 1992), Graded Ratings Scale Model (GRSM; Andrich, 1978), and Nominal Response Model (NRM; Bock, 1972). The test-level model-fit indices were used to compare and select IRT models, which included Akaike's information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwarz, 1978), and −2Log-Likelihood (−2LL; Spiegelhalter et al., 1998). The smaller values of these indices showed the better model fit; therefore, the model with the smallest test-fit indices was selected for further analysis. Model selection analysis was done in R package mirt (Version 1.10; Chalmers, 2012).

Step 3: Assess the local independence of the remaining items in the item pool.

Local independence includes two aspects: one is that the response of the same participants (or similar-level participants) to any one item will not be affected by any other items on the same test; and the other is that the responses of different participants (or different-level participants) on the same item do not affect each other (Embretson and Reise, 2013). Currently, the Q3 statistic (Yen, 1993) is commonly used to verify the dependent relationship between items. We calculated the Q3 values of any two items from the item pool under the selected IRT model in Step 2, via R package mirt (Version 1.10; Chalmers, 2012). As suggested by Cohen (2013), Q3 values below 0.36 represented local independence. Hence, one item with Q3 $>$ 0.36 in item pairs was removed.

Step 4: Assess the monotonicity of the remaining items in the item pool.

Monotonicity, meaning that a person with higher latent trait levels raises the possibility of higher scores for an item, was assessed by scalability coefficients for the item pool and individual items via R package Mokken (Version 2.7.7; van der Ark, 2007). According to Mokken (1971), a scale or item has high quality if the scalability coefficient is above 0.3. Items with scalability coefficients below 0.30 were thus eliminated until all of the scalability coefficients exceeded 0.3.

Step 5: Analyze the psychometric characteristics of the remaining items in item pool.

After items were excluded in the above four steps, psychometric characteristics (i.e., item-fit, differential item functioning [DIF], and discrimination) were evaluated for the remaining items. First, the S-X^2 statistic (Orlando and Thissen, 2003) was used to exam item fit using R package mirt (Version 1.10; Chalmers, 2012). Second, ordinal logistic regression, a nimbler method in detecting DIF, was used to test DIF for gender (male and female), age (under 18 years, and 18 and above), and region groups (rural and urban), respectively, via R package lordif (Version 0.2-2; Choi et al., 2011). DIF was assessed by means of change in McFadden's R^2 between different groups; items with R^2 change greater than 0.02 indicated DIF (Choi et al., 2011). The item parameters, namely, the discrimination (a) and difficulty parameters (b), were estimated under the selected model.

Step 6: Choose high-quality items to develop the final item bank of CAT-IA.

According to the psychometric characteristics in Step 5, poor model-fit ($p < 0.01$), DIF, and low discrimination items ($a < 1.00$) were all excluded. This procedure was repeated until no item was excluded.

CAT Simulation

To evaluate the psychometric properties, efficiency, reliability, concurrent validity, and predictive validity of CAT-IA, two CAT simulation studies were carried out. A CAT study is generally composed of six parts: the item bank, item response models, selection methods of initial items, evaluation methods of latent trait, item selection methods, and the stopping rules (Weiss and Kingsbury, 1984). First, the 59-item bank of CAT-IA was established, and the item parameters were estimated under the selected IRT model. Second, an item from the 59-item bank was randomly selected as the initial item to control the exposure rate. Ability estimation methods mainly include maximum likelihood estimation (MLE), weighted likelihood estimation (WLE), maximum a posteriori estimation (MAP), and expected a posterior estimation (EAP) in CAT procedures (e.g., Chen et al., 1998; Wang and Vispoel, 1998; Gorin et al., 2005). The MAP, MLE, and EAP methods regard the maximum point of the likelihood function (or posterior distribution) as the estimated ability value, which may result in multiple extreme points at the beginning of tests (Magis and Raîche, 2010). However, the mean value of the whole posterior distribution is adopted in EAP algorithm. Thus, the information provided by the entire posterior distribution can be effectively utilized, and the stability of the EAP algorithm is higher than that of the other three methods. The EAP method uses the mean value of the entire posterior distribution; therefore, it need not be iterated, and the calculation process is simpler. Compared with the MLE and WLE methods, the EAP method has a larger bias and belongs to biased estimation (Wang et al., 1999). Compared with the EAP method, the main advantage of MAP is that it requires fewer items in the variable-length test, which means that the test is more efficient (Wang and Vispoel, 1998). However, the virtues of the EAP algorithm importantly overweigh its drawbacks. The simplicity and stability of the EAP method makes it an optimal method for CAT simulations (e.g., Warm, 1989; Chen et al., 1998; Bulut and Kan, 2012). Further, maximum information criterion (MIC; Lord, 1980) is the most widely used item selection strategy in CAT programs because of its relatively simple implementation method. The purpose of this strategy is to improve the accuracy of measurement (Brunel and Nadal, 1998), but it can easily lead to uneven exposure of items in the item bank and reduced security of the test (Barrada et al., 2008). Different from the exam, a Likert-type scale without correct answers requires participants to respond in the usual way, which greatly reduces the test security problem. Therefore, MIC was selected as the item selection method in the CAT-IA simulation study. Finally, several stopping rules with different SEs were performed, including None (i.e., the entire item bank was used), SE ≤ 0.2, SE ≤ 0.3, SE ≤ 0.4, and SE ≤ 0.5, respectively.

Simulation Study 1: Psychometric Properties of CAT-IA

When a CAT-IA program is established, its psychometric properties should be evaluated, especially in terms of measurement accuracy. The results of CAT-IA may result in high-risk outcomes that are similar to the entrance exam. Therefore, the Monte-Carlo (MC) simulation method was used to evaluate the performance of CAT-IA. First, the ability of 1,000 virtual persons were generated randomly from the normal distribution (Mean = 0, SD = 1); this sample was regarded as the true ability values. Second, the item parameters of the final 59-item bank and selected IRT model were used to conduct the CAT-IA simulation study. Third, the MC method was used to estimate the ability value of each participant according to the true θ values, selected IRT model and item parameters. These abilities were the estimated values of 1,000 simulated persons. In addition, the CAT-IA performance was evaluated via several statistical indices, including conditional bias (CBIAS), conditional mean absolute error (CMAE), conditional root mean square error (CRMSE), and conditional standard error of estimation (CSEE) across all θ areas (Han, 2018). Simulation study 1 was done in the R package catR (Version 3.12; Magis and Barrada, 2017). These statistical indices for every participant were plotted under different stopping rules using SPSS (Version 23.0; George, 2016).

Simulation Study 2: Efficiency, Reliability, and Validity of CAT-IA

Efficiency and reliability of CAT-IA

To evaluate the efficiency and reliability of CAT-IA, a simulation based on the actual data was carried out via the R package mirtCAT (Version 0.5; Chalmers, 2015). In simulation study 2, the real responses to items were used instead of virtual responses generated by the MC method; the process of simulation study 2 was the same as that in simulation study 1. For each responder, the SE could be calculated in simulation study 2. Green et al. (1984) pointed out that a unitless reliability index is necessary for a CAT, even if this index is somewhat contrived. The index of marginal reliability was proposed by Green et al. (1984) to evaluate effectively the reliability of a CAT under different stopping rules. Marginal reliability is a relatively convenient way to monitor dynamically the reliability of a CAT, and can also be used to evaluate the stability of a CAT (Green et al., 1984). In general, marginal reliability is a function of standard error of measurement (SEM), as shown in formulas (1) and (2). The bigger the marginal reliability is, the smaller the SEM is. Therefore, marginal reliability is crucial for the assessment of SEM and the reliability of measurement in CAT. Marginal reliability is equal to the mean reliability under each stopping rule for all participants (Wainer et al., 2000b). The formula of marginal reliability is defined as:

$$MR = 1 - SE^2 \tag{1}$$

$$SE = \frac{\sum_{i=1}^{N} SE(\theta_i)}{N} \tag{2}$$

Where n is the number of all participants, and $SE(\theta_i)$ is the standard error of examinee i at the finally estimated θ. Some statistics were investigated to examine the efficiency and reliability of CAT-IA, including the mean and standard deviation of the used items, mean SE, marginal reliability, and Pearson's correlations between the estimated θ with the stopping rule of None and the remaining stopping rules. The number of used items with the reliability for every participant was plotted under different stopping rules using the R package ggplot2 (Version 2.2.1; Wickham, 2011).

Concurrent validity and predictive validity of CAT-IA

CAT-IA may take effect when CAT-IA estimation results have a favorable similarity to the results of the existing widely used scales. In other words, a person who is diagnosed with IA in a questionnaire has a higher latent trait in a CAT estimation compared with those without a diagnosis of IA. The similarities were evaluated by concurrent validity and predictive validity of CAT-IA using SPSS (Version 23.0; George, 2016) based on the initial responses that were used to establish the item bank of IA. The concurrent validity was evaluated by the Pearson's correlations between the estimated θ of CAT-IA and the aggregate scores of each scale. Based on previous studies, only two scales (IAT and GAS) possess the definite diagnostic criteria for IA (Young, 1998; Caplan, 2002; Huang et al., 2007; Lemmens et al., 2009). Individuals whose sum scale scores of IAT exceed 39 are considered as having problematic network usage (Young, 1998). GAS includes seven diagnostic items (Lemmens et al., 2009); individuals with at least four items scoring 4 or 5 are considered to be addicted. The diagnostic results of IAT and GAS were used to compare the estimated results of CAT-IA. Then, the AUC (the area under ROC curve) index was employed to investigate the predictive effect of CAT-IA. According to the rule of Rice and Harris (2005), AUC values below 0.50 represent a small predictive effect; values between 0.51 and 0.70, a moderate predictive effect; and values higher than 0.71, a large predictive effect. In the ROC curve, determination of the critical points adopted the maximal Youden Index (YI = sensitivity + specificity − 1) (Schisterman et al., 2005). The sensitivity indicates the probability of a patient being diagnosed as a patient, and the specificity indicates the probability of a person without the symptoms being diagnosed as a normal person. Sensitivity and specificity are two important reference indicators for the accuracy of critical values, which are both ranged from 0 to 1, with the bigger values representing better predictive validation.

RESULTS

Item Bank Construction of CAT-IA

Unidimensionality

In EFA, the ratio of variance explained by the first factor was 32.44% higher than the critical standard of 20% (Reckase, 1979), and the ratio of variance explained in the first and second factors was 5.89 higher than the critical standard of 4 (Reeve et al., 2007). In the single-factor CFA, five items were removed (see **Table 3**) owing to their factor loadings of below 0.4 (Nunnally, 1978). Both the EFA and single-factor CFA were again conducted on the remaining 96 items. The EFA results showed the ratio of variance explained by the first factor was 33.87%, and the ratio of variance explained in the first and second factors was 6.14. Results of the single-factor CFA indicated that the RMSEA value was 0.08, indicating that the single factor model was fair or acceptable; all factor loadings were above 0.4. The above results showed that the remaining 96 items, after deleting five items, basically met the unidimensional hypothesis.

TABLE 3 | Reasons for stepwise exclusion of the items.

Excluded reasons	Excluded items
Unidimensionality	IAT-7 and 9, GPIUS-36 and 37 and CIAT-100
Local Independency	IAT-4 and 16; GPIUS-22, 23, 25, 26, 27, 28, 31, 39, 40, 42, and 48; GAS-50, 52, 51, 53, 54, 57, 58, 60, 62, and 63; CIAT-87, 89, and 90
Monotonicity	IAT-1 and 5; GPIUS-21, 30, and 47; CIAT-73
DIF	GAS-61, 64, 67, and 69
$S\text{-}X^2$	IAT-2
Discrimination	None

DIF, different item function; the abbreviated content of each item can be seen in ***Table 5****.*

Model Selection

Table 4 documents the model-fit indices, including −2LL, AIC, and BIC, for the four IRT models. Compared with the other three IRT models, the GRSM fitted the worst in that it had the largest −2LL, AIC, and BIC values. Of the remaining three models, the GPCM model had the worst fitting indices. Although the −2LL value of NRM was smaller than that of GRM, the AIC and BIC values of NRM were both higher compared with the GRM. The GRM model overall fitted the remaining 96-item bank best compared with other three. Therefore, GRM was selected for later analysis.

Local Independence

A total of 23 pairs of items showed local dependence: their Q3 values were above 0.36 (Cohen, 2013). Thus, 26 items were excluded owing to local dependence, including 2 IAT items, 11 GPIUS items, 10 GAS items, and 3 CIAT items (see **Table 3**). Then, the Q3 values of the remaining 70-item bank were reassessed, and the results showed all Q3 values were below 0.36.

Monotonicity

The scalability coefficient for the remaining 70-item bank was 0.4, which was higher the requirement of 0.3 (Mokken, 1971). However, for the scalability coefficient of the 70 items, there were still six items (see **Table 3**) with scalability coefficients below 0.3. After excluding these items, we reevaluated the scalability coefficients, and the results showed that the scalability coefficient of the 64-item bank was 0.39, whereas all scalability coefficients of the 64 items were above 0.3.

DIF

For the region and age groups, no DIF was found for all 64 items; the means of change in McFadden's R^2 between different groups were above the minimum requirement of 0.02 (Choi et al., 2011). However, for the gender group, four items (see **Table 3**), all belonging to GAS, were flagged for DIF. Therefore, we excluded these items and reassessed the DIF of 60 items. The results

TABLE 4 | Model-fit indices.

Model	−2LL	AIC	BIC
GRM	331710.400	332670.500	335217.000
GPCM	333965.400	334925.300	337471.800
GRSM	336329.000	336719.000	337753.500
NRM	331675.600	333211.600	337286.000

GRM, Graded Response Model; GPCM, Generalized Partial Credit Model; GRSM, Graded Ratings Scale Model; NRM, Nominal Response Model; −2LL, −2Log-Likelihood; AIC, Akaike's information criterion; BIC, Bayesian information criterion.

TABLE 5 | Item parameters for 59-item bank with GRM.

Item	Abbreviated	*a*	*b1*	*b2*	*b3*	*b4*	Domain
IAT-3	Excitement	1.587	−0.540	1.054	2.425	3.092	Mood modification
IAT-6	Work suffer	1.369	−1.365	0.098	1.724	3.213	Negative outcomes
IAT-8	Job suffer	1.292	−1.352	0.071	1.723	3.351	Negative outcomes
IAT-10	Block disturbing	1.072	−1.545	0.057	2.049	3.772	Mood modification
IAT-11	Anticipating	1.236	−1.447	0.204	1.656	2.945	Tolerance
IAT-12	Boring and joyless	1.284	−1.344	−0.202	1.356	2.633	Withdrawal
IAT-13	Annoyed	1.473	−0.466	1.175	2.519	3.457	Withdrawal
IAT-14	Lose sleep	1.397	−0.932	0.499	1.846	3.331	Negative outcomes
IAT-15	Preoccupied	1.863	−0.521	0.797	2.14	3.144	Salience
IAT-17	Fail to reduce time	1.474	−1.258	0.027	1.415	2.648	Relapse
IAT-18	Hide online time	1.302	−0.238	1.255	2.694	3.956	Negative outcomes
IAT-19	Prefer online	1.630	−0.190	1.079	2.316	3.149	Salience
IAT-20	Depressed or nervous	1.972	−0.183	1.133	2.298	3.202	Withdrawal
GPIUS-24	Feel better	1.352	−1.352	−0.303	0.838	2.852	Mood modification
GPIUS-29	Treated better	1.324	−0.725	0.600	1.916	3.438	Benefits
GPIUS-32	Feel worthless offline	1.397	0.217	1.532	2.541	4.121	Benefits
GPIUS-33	Missed social event	1.298	−0.274	0.975	2.092	3.320	Negative outcomes
GPIUS-34	Unsuccessful	1.602	−0.829	0.277	1.209	2.686	Relapse
GPIUS-35	Fail to reduce time	1.631	−0.659	0.452	1.495	2.661	Relapse
GPIUS-38	Forget the time	1.101	−1.525	−0.367	0.595	2.595	Tolerance
GPIUS-41	Longer time	1.404	−1.477	−0.382	0.495	2.531	Tolerance
GPIUS-43	Miss	1.657	−0.976	0.078	0.939	2.545	Withdrawal
GPIUS-44	Wonder	1.335	−1.23	−0.074	0.836	2.867	Withdrawal
GPIUS-45	Feel lost	1.856	−0.675	0.358	1.225	2.696	Withdrawal
GPIUS-46	Unable to stop thinking	1.659	−0.578	0.516	1.481	2.785	Tolerance
GPIUS-49	Control	1.247	−0.400	0.858	2.204	3.795	Benefits
GAS-55	Unable to stop playing	1.381	−0.418	0.848	2.147	3.081	Tolerance
GAS-56	Forget about real life	1.534	−0.099	1.211	2.602	3.401	Mood modification
GAS-59	Unable to reduce time	1.490	−0.207	1.163	2.294	3.205	Relapse
GAS-65	Fights with others	1.719	−0.173	0.960	2.179	3.142	Negative outcomes
GAS-66	Neglected others	1.787	−0.365	0.635	1.967	2.995	Negative outcomes
GAS-68	Lose sleep	1.721	−0.32	0.777	1.948	2.795	Negative outcomes
GAS-70	Feel bad	1.195	−1.299	−0.119	1.611	3.131	Negative outcomes
CIAT-71	Neglect household	1.984	−0.722	0.462	1.687	2.714	Negative outcomes
CIAT-72	Excitement	2.294	−0.374	0.794	1.806	2.629	Mood modification
CIAT-74	Complain of others	1.745	−0.449	0.906	2.061	2.981	Negative outcomes
CIAT-75	School or work suffer	1.879	−0.848	0.334	1.544	2.628	Negative outcomes
CIAT-76	Defensive or secretive	1.189	−0.845	0.767	2.322	3.384	Negative outcomes
CIAT-77	Disturbing	1.631	−1.006	0.152	1.614	2.733	Mood modification
CIAT-78	Anticipating	1.975	−0.742	0.445	1.695	2.513	Tolerance
CIAT-79	Annoyed act	1.831	−0.176	1.157	2.132	2.994	Withdrawal
CIAT-80	Lose sleep	1.456	−0.498	0.739	1.975	2.876	Negative outcomes
CIAT-81	Preoccupied	2.639	−0.404	0.728	1.782	2.407	Salience
CIAT-82	"Just a few minutes"	2.053	−0.866	0.206	1.354	2.421	Relapse
CIAT-83	Hide online time	1.873	−0.207	1.095	2.122	3.108	Negative outcomes
CIAT-84	Spend more time	2.409	−0.375	0.556	1.429	2.242	Tolerance
CIAT-85	Important	2.077	−0.343	0.631	1.63	2.521	Salience
CIAT-86	More attractive	2.093	−0.337	0.744	1.904	2.795	Benefits
CIAT-88	Exciting information	1.382	−1.450	−0.112	1.581	2.845	Benefits
CIAT-91	Reduce the stress	1.443	−1.298	−0.006	1.659	2.918	Benefits
CIAT-92	Times goes faster	1.189	−1.968	−0.854	0.511	2.143	Tolerance
CIAT-93	Stay online	2.192	−0.787	0.404	1.380	2.349	Tolerance
CIAT-94	Want to stay online	2.233	−0.825	0.421	1.605	2.350	Withdrawal

(Continued)

TABLE 5 | Continued

Item	Abbreviated	*a*	*b1*	*b2*	*b3*	*b4*	Domain
CIAT-95	Disturbed	1.219	−1.951	−0.632	0.800	2.250	Withdrawal
CIAT-96	Distraught	1.894	−0.713	0.511	1.622	2.621	Withdrawal
CIAT-97	Failed to reduce time	2.103	−0.698	0.493	1.575	2.443	Relapse
CIAT-98	Addiction	1.391	−1.158	−0.098	1.259	2.499	Salience
CIAT-99	Addiction	1.675	−0.579	0.721	1.796	2.730	Salience
CIAT-101	Dependent	1.504	−0.308	1.141	2.388	3.182	Relapse

a, discrimination parameter; b, difficulty parameter.

TABLE 6 | The psychometric properties of CAT-IA using CBIAS, CMAE, CRMSE, and CSEE indices across all θ areas.

Stopping rule	CSEE	CBIAS	CMAE	CRMSE
None	0.154	−0.005	0.125	0.160
SE (θ) ≤ 0.2	0.200	0.003	0.158	0.199
SE (θ) ≤ 0.3	0.292	0.007	0.227	0.283
SE (θ) ≤ 0.4	0.380	0.008	0.278	0.348
SE (θ) ≤ 0.5	0.464	−0.016	0.359	0.456

None, all item bank was used; CBIAS, conditional bias; CMAE, conditional mean absolute error; CRMSE, conditional root mean square error; CSEE, conditional standard error of estimation.

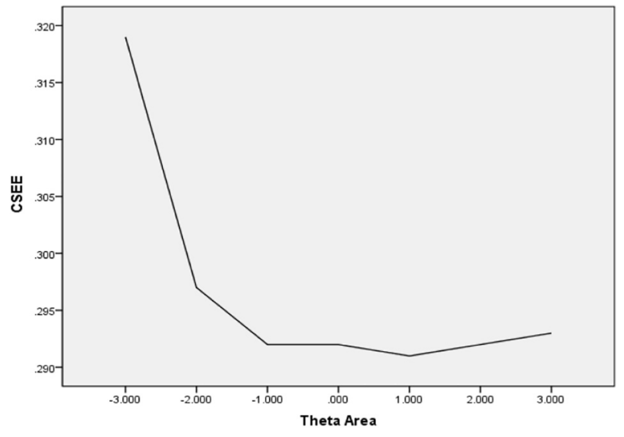


FIGURE 1 | Conditional SEE (average SEE in each theta area).

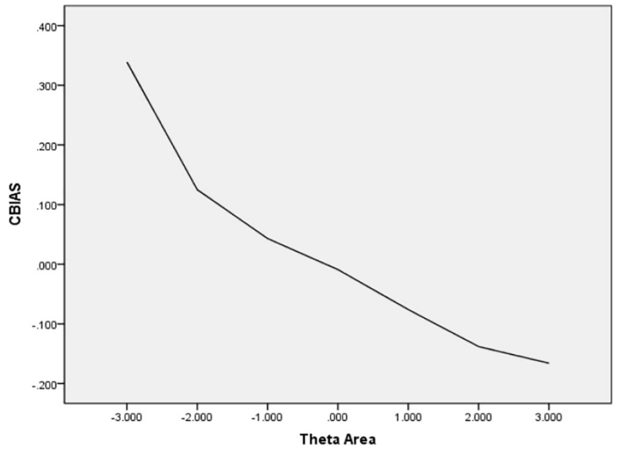


FIGURE 2 | Conditional BIAS (average BIAS in each theta area).

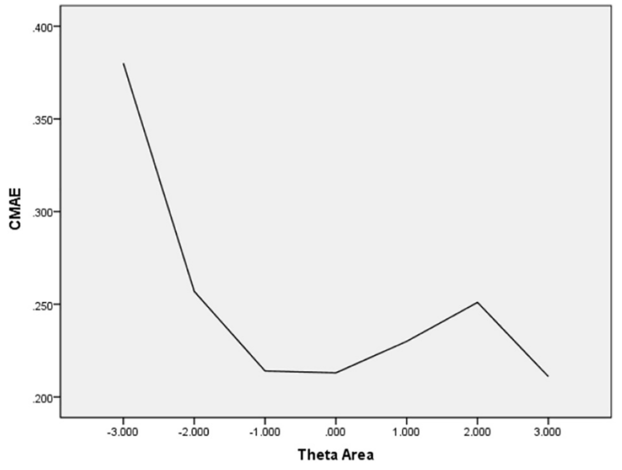


FIGURE 3 | Conditional MAE (average MAE in each theta area).

showed that the means of change in McFadden's R^2 all were below 0.02 for the region, age, and gender groups.

Item-Fit

Only one item (IAT-2) failed to fit the GRM for having a *p*-value of $S{-}X^2$ that was less than 0.01. After removing this item, the remaining 59 items were reevaluated, and the results showed that the *p*-value of $S\text{-}X^2$ of all the 59 items were above 0.01.

Discrimination

Graded Response Model was used again to calibrate the remaining 59 items. The item parameters are listed in **Table 5**. The discrimination parameters of the 59 items were all above the value of 1 with mean of 1.627 (*SD* = 14.5), which indicated the final item bank was of a high quality.

After the above steps, the final item bank of CAT-IA included 59 items with high discrimination, good item-fit, no DIF, and meeting the assumptions of IRT. The eighth column in **Table 5** shows the domains of the 59 items: 6 items measured salience, 9 items measured tolerance, 6 items measured mood

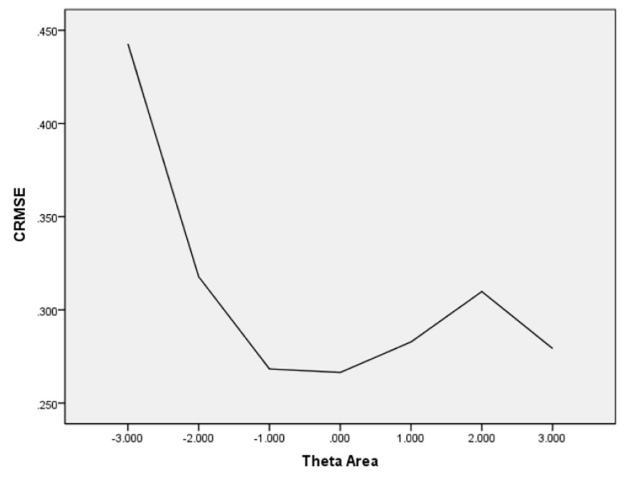


FIGURE 4 | Conditional RMSE (average RMSE in each theta area).

modification, 7 items measured relapse, 10 items measured withdrawal, 16 items measured negative outcomes, and 6 items measured benefits.

Psychometric Properties of CAT-IA

In **Table 6**, the values of CBIAS, CMAE, CRMSE, and CSEE across all θ areas are displayed under several stopping rules. The second column documents the CSEE values across all θ areas, which ranged from 0.154 to 0.464. The values of CSEE across all θ areas that were less than the corresponding measurement precision decreased as measurement precision was made stricter. The third column reveals the values of CBIAS across all θ areas, which ranged from −0.016 to 0.008. Except for the stopping rule of SE (θ) ≤ 0.5, with CBIAS of −0.016 across all θ areas, the values of CBIAS across all θ areas decreased when the measurement precision was made stricter. The last two columns of **Table 6** indicate that the CMAE and CRMSE values across all θ areas varied from 0.125 to 0.359, and 0.160 to 0.456, respectively. The values of CMAE and CRMSE across all θ areas decreased as measurement precision was made stricter, respectively. All these results indicated that the CAT-IA had high measurement accuracy in psychometric properties. The values of CBIAS, CMAE, CRMSE, and CSEE in each θ area under stopping rule SE (θ) ≤ 0.3 are displayed in **Figures 1–4**. Clearly, as shown in **Figure 1**, the CSEE values were closely commanded to less than 0.3 at −2 ≤ θ area. The values of CBIAS were inversely proportional to all θ areas. In addition, CBIAS values gradually decreased as the ability increased, as shown in **Figure 2**. The changing trends of CMAE and CRMSE were approximately consistent across all θ areas, as shown in **Figures 3**, **4**. These results were consistent for all stopping rules.

Efficiency, Reliability, and Validity of CAT-IA

Efficiency and Reliability of CAT-IA

In **Table 7**, the CAT-IA simulation results are displayed under five measurement precision standards. As shown in the second column, the mean and SD of the items used both increased when the measurement precision was made stricter. In the third column, the mean SE of the latent traits for each stopping rule varied from 0.159 to 0.454. Except for the stopping rule of SE (θ) ≤ 0.2, the mean SEs were less than their corresponding measurement precision. Marginal reliability ranged from 0.794 to 0.973 with an average of 0.90, as shown in the fourth column. Evidently, marginal reliability increased as the measurement precision was made stricter. The last column in **Table 7** shows the Pearson's correlation between the estimated θ with stopping rule of None and the remaining stopping rules. The values of Pearson's correlation ranged from 0.898 to 1 and were all significant at the 0.01 level (two-tailed), which showed that under different stopping rules, the algorithm of CAT-IA was effective. **Table 7** also shows that the CAT-IA could greatly save item usage without loss of measurement precision. Under the stopping rule of SE (θ) ≤ 0.2, the Pearson's correlation between the estimated theta by CAT-IA and the estimated theta by all of the items in the item bank reached 0.990; CAT-IA only used about half of the items (27.655 items) in the item bank. In brief, the CAT-IA saved 53.1% in item usage without loss of measurement precision. Under the two stopping rules of SE (θ) ≤ 0.3 and SE (θ) ≤ 0.4, the Pearson's correlations were both above 0.90; CAT-IA thus saved 80.7 and 89.9% of item usage, respectively. All these results indicated that the CAT-IA had high efficiency and marginal reliability.

The reliability and number of used items in CAT-IA on levels of the latent trait under stopping rule SE (θ) ≤ 0.3 are displayed in **Figure 5**. We noted a remarkable connection between the number of used items and reliability. Despite only using about 11.38 items, the CAT-IA obtained high reliability (above 0.9) and

TABLE 7 | CAT simulation statistics for CAT-IA under different stopping rules.

Stopping rule	Number of items used		Mean SE (theta)	Marginal reliability	*r*
	Mean	*SD*			
None	59	0	0.159	0.975	1
SE (θ) ≤ 0.2	27.655	12.070	0.203	0.959	0.990**
SE (θ) ≤ 0.3	11.380	9.064	0.293	0.914	0.962**
SE (θ) ≤ 0.4	5.952	4.819	0.380	0.856	0.932**
SE (θ) ≤ 0.5	3.675	2.000	0.454	0.794	0.898**

*None, all item bank was used; SD, standard deviation; SE, standard error; r, Pearson's correlations. ** representing significant correlation at the 0.01 level (two-tailed).*

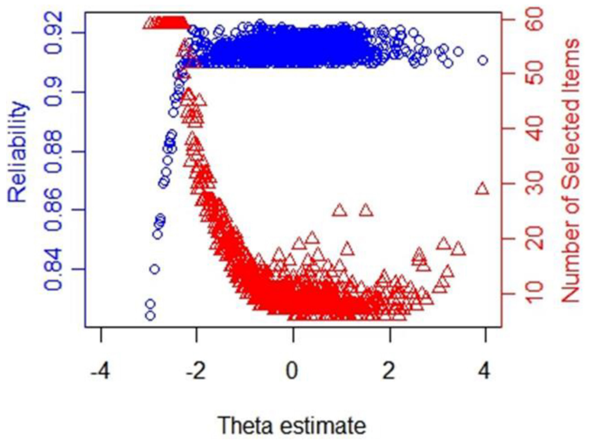


FIGURE 5 | Number of selected items and reliability shown as a function of the final θ estimate under stopping rule SE (θ) ≤ 0.3.

TABLE 8 | Pearson's correlations between the estimated θ of CAT-IA and the sum scores of four IA scales under different stopping rules.

Stopping rules	IAT	GPIUS	GAS	CIAT
None	0.862**	0.861**	0.754**	0.944**
SE (θ) ≤ 0.2	0.825**	0.839**	0.731**	0.941**
SE (θ) ≤ 0.3	0.781**	0.796**	0.684**	0.926**
SE (θ) ≤ 0.4	0.757**	0.773**	0.669**	0.893**
SE (θ) ≤ 0.5	0.728**	0.740**	0.646**	0.858**

*None, all item bank was used; IAT, Internet Addiction Test; GPIUS, Generalized Problematic Internet Use Scale; GAS, Gaming Addiction Scale; CIAT, Chinese Internet Addiction Test; ** representing significant correlation at the 0.01 level (two-tailed).*

high measurement precision for a large number of individuals (estimated theta ranged from −2 to +4). Conversely, when the reliability was below 0.9, more items were used. This result was consistent for all stopping rules.

Concurrent Validity and Predictive Validity of CAT-IA

The Pearson's correlations between the estimated θ of CAT-IA and the aggregate scores of IAT, GPIUS, GAS, and CIAT are documented in **Table 8**. The values of Pearson's correlations varied from 0.646 to 0.944 and were all significant at the 0.01 level (two-tailed), which revealed that the CAT-IA had high concurrent validity. In addition, comparing the other scales, the correlation coefficient of CIAT was the highest under each stopping rule, whereas that of GAS was the lowest.

The results of the predictive validity of CAT-IA are displayed in **Table 9**. All AUC values (with 95% confidence intervals) were above 0.71, indicating that CAT-IA had a large predictive effect (Rice and Harris, 2005). According to the large predictive effect, the cut-off point of IA was determined under each stopping rule for IAT and GAS, based on the values of sensitivity and specificity. For example, under the stopping rule of SE (θ) ≤ 0.2 in the diagnostic criteria of GAS, if the cut-off point of the 59-item bank was set to 0.801, the sensitivity and specificity of CAT-IA reached 0.922 and 0.862, respectively. These results showed that the CAT-IA had high predictive validity and had strong discrimination between individuals with IA disorder and healthy individuals.

DISCUSSION

CAT studies have focused on depression or anxiety for clinical individuals in the field of mental health (e.g., Fliege et al., 2005; Flens et al., 2016, 2017). However, to the best of our knowledge, there are no CAT studies on IA. In this research, we developed a CAT-IA to provide a new and effective assessment of IA. The original item bank of IA was subjected to psychometric evaluation; items were excluded until all of the remaining items in the item bank satisfied the requirements of psychometric evaluation. Subsequently, the efficiency, reliability, and validity of the final item bank of the CAT-IA were assessed under different stopping rules. The results showed that the final 59-item CAT-IA item bank met the three IRT assumptions, and possessed high discrimination, good item-model fit, and no DIF. Moreover, the CAT-IA could significantly save testing items and effectively reduce the test burden of participants, while also having high reliability, concurrent validity, and predictive validity.

Kocalevent et al. (2009) demonstrated that simulation and actual results of CAT tend to show high similarity. There are three reasons to implement actual CAT studies under different stopping rules. First, the same participants are used not only to estimate item parameters but also to simulate CAT studies, which could result in overfitting and more optimistic results (Friedman et al., 2010). Second, margin reliability and predictive validity might be overestimated because the data of CAT simulation

TABLE 9 | Area under the curve Statistics for the IAT and GAS scale under different stopping rules, and 95% confidence intervals.

Stopping rules	GAS				IAT			
	AUC [95% CI]	Cut-off	Se	Sp	AUC [95% CI]	Cut-off	Se	Sp
None	0.957 [0.933, 0.981]	0.750	0.969	0.838	0.931 [0.913, 0.950]	0.749	0.815	0.882
SE (θ) ≤ 0.2	0.948 [0.921, 0.976]	0.801	0.922	0.862	0.903 [0.887, 0.918]	0.203	0.865	0.773
SE (θ) ≤ 0.3	0.927 [0.895, 0.958]	0.946	0.813	0.893	0.875 [0.856, 0.893]	0.205	0.825	0.746
SE (θ) ≤ 0.4	0.919 [0.884, 0.954]	0.868	0.828	0.873	0.863 [0.844, 0.882]	0.151	0.835	0.728
SE (θ) ≤ 0.5	0.906 [0.868, 0.944]	0.780	0.797	0.860	0.848 [0.828, 0.868]	0.088	0.861	0.673

None, all item bank was used; IAT, Internet Addiction Test; GAS, Gaming Addiction Scale. Se, sensitivity; Sp, specificity.

studies come from the original database. Third, De Beurs et al. (2012) indicated that the results of a test are affected by the measurement tools. The original CAT study was done on a computer, but now it is conducted as a paper-and-pencil survey, which may lead to different outcomes.

When applying CAT-IA in clinical practice or research, CAT-IA may have different reliability results for different observers; that is, individuals of different abilities are provided with different information. For example, in the present study, under the stopping rule SE (θ) $\leq$ 0.3, reliability was very low and a large number of items were used when the individual has overly high or low abilities, indicating that small differences between two participants with either very high or very low abilities may not be detected, which was similar to Reise and Waller (2009) findings. To prevent the emergence of test bias, the reliability provided by the CAT-IA was set as similar and high for all test-takers. Nonetheless, we recognized the impact of the difficulty parameter distribution under the GRM. For example, in this study, there were no items to match persons whose abilities are below −1.968 in that the minimum value of the difficulty parameters was b1 = −1.968. Therefore, the CAT-IA provided these people with scarce information, and the measurement accuracy and reliability for them were very low despite the use of a large number of items of the 59-item bank. In future studies, researchers can increase the number of items with high or low difficulty parameter to make the difficulty parameter reasonable, which could not only provide high measurement accuracy and reliability for each participant but also greatly reduce the number of selected items for each person.

The standard IRT model is generally based on assumptions of unidimensionality and local independence. However, the single-dimensional and locally independent assumptions in real life may not be completely satisfied. For example, many researchers believe that the factor structure of IA should be multidimensional rather than unidimensional (e.g., Thatcher and Goolam, 2005; Lemmens et al., 2009; Caplan, 2010). Based on local dependency, Wainer et al. (2000a) proposed a widely used 3PL testlet model, in which dependent items did not need to be excluded when the testlet model was used in a CAT. According to these results, future studies can extend the unidimensional CAT into the multidimensional CAT and use the testlet model to solve local dependency between items.

In addition, concurrent validity in the present study was evaluated by Pearson's correlations between the estimated θ. of CAT-IA and the aggregate scores of each scale. This method can result in item overlap that may overestimate the concurrent validity. Future studies should utilize other external scales to investigate concurrent validity. Further, De Beurs et al. (2012) proved that the same test applied in different situations may lead to changes in the measurement characteristics. Therefore, factorial invariance should be considered in future research. Lastly, although there are many methods for the selection of initial items, with respect to the estimation of latent trait, item selection, and exposure rate, this study failed to address enough methods (such as different parameter estimation and item selection methods), which should be fully considered in future studies.

ETHICS STATEMENT

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study. The current study was conducted in conformity to the recommendations of psychometrics studies on mental health at the Research Center of Mental Health, Jiangxi Normal University and approved by the Research Center of Mental Health, Jiangxi Normal University and the Ethics Committee of Psychology Department in Jiangxi Normal University. The written informed consent was obtained from all participants in accordance with the Declaration of Helsinki. All participants gave their written informed consent. The parental consent was also obtained for all participants under the age of 16.

AUTHOR CONTRIBUTIONS

YZ wrote the manuscript. YC and DT guided the manuscript writing and data processing. DW and XG processed the data.

FUNDING

This study was funded by the National Natural Science Foundation of China (Grant Nos. 31760288 and 31660278).

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 716–723. doi: 10.1109/tac.1974.1100705

Almond, R. G., and Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Appl. Psychol. Meas.* 23, 223–237. doi: 10.1177/0146621615590401

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika* 43, 561–573. doi: 10.1007/bf02293814

Barrada, J. R., Olea, J., Ponsoda, V., and Abad, F. J. (2008). Incorporating randomness in the fisher information for improving item-exposure control in CATs. *Br. J. Math. Stat. Psychol.* 61, 493–513. doi: 10.1348/000711007x230937

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* 37, 29–51. doi: 10.1007/bf02291411

Browne, M. W., and Cudeck, R. (1993). "Alternative ways of assessing model fit," in *Testing Structural Equation Models*, eds K. A. Bollen and J. S. Long. Newbury Park, CA: Sage

Brunel, N., and Nadal, J. P. (1998). Mutual information, fisher information, and population coding. *Neural Comput.* 10, 1731–1757. doi: 10.1162/089976698300017115

Bulut, O., and Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Eurasian J. Educ. Res.* 49, 61–80.

Caplan, S. E. (2002). Problematic internet use and psychosocial well-being: development of a theory-based cognitive–behavioral measurement instrument. *Comput. Hum. Behav.* 18, 553–575. doi: 10.1016/s0747-5632(02)00004-3

Caplan, S. E. (2010). Theory and measurement of generalized problematic internet use: a two-step approach. *Comput. Hum. Behav.* 26, 1089–1097. doi: 10.1016/j.chb.2010.03.012

Chalmers, R. P. (2012). Mirt: a multidimensional item response theory package for the R environment. *J. Stat. Softw.* 48, 1–29.

Chalmers, R. P. (2015). *mirtCAT: Computerized Adaptive Testing With Multidimensional Item Response Theory. R Package Version 0.6.*

Chen, S. K., Hou, L., and Dodd, B. G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educ. Psychol. Meas.* 58, 569–595. doi: 10.1177/0013164498058004002

Choi, S. W., Gibbons, L. E., and Crane, P. K. (2011). Lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *J. Stat. Softw.* 39, 1–30.

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Abingdon: Routledge.

De Beurs, E., Barendregt, M., Flens, G., van Dijk, E., Huijbrechts, I., and Meerding, W. J. (2012). Equivalentie in responsiviteit van veel gebruikte zelfrapportage meet instrumenten in de geestelijke gezondheidszorg [Equivalence in responsiveness of commonly used self-report questionnaires in mental health]. *Maandblad voor de Geestelijke Volksgezondheid* 67, 259–264.

Embretson, S. E., and Reise, S. P. (2013). *Item Response Theory*. Hove: Psychology Press.

Flens, G., Smits, N., Carlier, I., van Hemert, A. M., and de Beurs, E. (2016). Simulating computer adaptive testing with the mood and anxiety symptom questionnaire. *Psychol. Assess.* 28:953. doi: 10.1037/pas0000240

Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., and de Beurs, E. (2017). Development of a computer adaptive test for depression based on the dutch-flemish version of the PROMIS item bank. *Eval. Health Prof.* 40, 79–105. doi: 10.1177/0163278716684168

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., and Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Qual. Life Res.* 14, 2277–2291. doi: 10.1007/s11136-005-6651-9

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22.

George, D. (2016). *IBM SPSS Statistics 23 Step by Step: A Simple Guide and Reference*. New York, NY: Routledge.

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., and Shieh, Y. Y. (2005). Computerized adaptive testing with the partial credit model: estimation procedures, population distributions, and item pool characteristics. *Appl. Psychol. Meas.* 29, 433–456. doi: 10.1177/0146621605280072

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., and Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *J. Educ. Meas.* 21, 347–360. doi: 10.1111/j.1745-3984.1984.tb01039.x

Han, K. C. T. (2018). Conducting simulation studies for computerized adaptive testing using SimulCAT: an instructional piece. *J. Educ. Eval. Health Prof.* 15:20. doi: 10.3352/jeehp.2018.15.20

Ho, R. C., Zhang, M. W., Tsang, T. Y., Toh, A. H., Pan, F., Lu, Y., et al. (2014). The association between internet addiction and psychiatric co-morbidity: a meta-analysis. *BMC Psychiatry* 14:183. doi: 10.1186/1471-244X-14-183

Huang, Z., Wang, M., Qian, M., Zhong, J., and Tao, R. (2007). Chinese internet addiction inventory: developing a measure of problematic internet use for Chinese college students. *Cyberpsychol. Behav.* 10, 805–812. doi: 10.1089/cpb.2007.9950

International Telecommunication Union (2016). *ICT Facts and Figures 2016*. Available at: https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2016pdf (accessed February 2018).

Ireland, C. M. (1977). *An Application of the Rasch one Parameter Logistic Model to Individual Intelligence Testing in a Tailored Testing Environment*. Ph.D. thesis, ProQuest Information & Learning, Michigan.

Kocalevent, R. D., Rose, M., Becker, J., Walter, O. B., Fliege, H., Bjorner, J. B., et al. (2009). An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception. *J. Clin. Epidemiol.* 62, 278–287. doi: 10.1016/j.jclinepi.2008.03.003

Lai, C. M., Mak, K. K., Watanabe, H., Ang, R. P., Pang, J. S., and Ho, R. C. (2013). Psychometric properties of the internet addiction test in Chinese adolescents. *J. Pediatr. Psychol.* 38, 794–807. doi: 10.1093/jpepsy/jst022

Lemmens, J. S., Valkenburg, P. M., and Peter, J. (2009). Development and validation of a game addiction scale for adolescents. *Media Psychol.* 12, 77–95. doi: 10.1080/15213260802669458

Linacre, J. M. (2000). "Computer-adaptive testing: a methodology whose time has come," in *Development of Computerised Middle School Achievement Tests, MESA Research Memorandum*, Vol. 69, eds S. Chae, U. Kang, E. Jeon, and J. M. Linacre (Seoul: Komesa Press).

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Magis, D., and Barrada, J. R. (2017). Computerized adaptive testing with R: recent updates of the package catR. *J. Stat. Softw.* 76, 1–19.

Magis, D., and Raîche, G. (2010). An iterative maximum a posteriori estimation of proficiency level to detect multiple local likelihood maxima. *Appl. Psychol. Meas.* 34, 75–89. doi: 10.1177/0146621609336540

Mak, K. K., Lai, C. M., Ko, C. H., Chou, C., Kim, D. I., Watanabe, H., et al. (2014a). Psychometric properties of the revised chen internet addiction scale (CIAS-R) in Chinese adolescents. *J. Abnorm. Child Psychol.* 42, 1237–1245. doi: 10.1007/s10802-014-9851-3

Mak, K. K., Lai, C. M., Watanabe, H., Kim, D. I., Bahar, N., Ramos, M., et al. (2014b). Epidemiology of internet behaviors and addiction among adolescents in six Asian countries. *Cyberpsychol. Behav. Soc. Netw.* 17, 720–728. doi: 10.1089/cyber.2014.0139

Mokken, R. J. (1971). *A Theory and Procedure of Scale Analysis: With Applications in Political Research*, Vol. 1. Berlin: Walter de Gruyter.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Appl. Psychol. Meas.* 16, 159–176. doi: 10.1177/014662169201600206

Muthén, L. K., and Muthén, B. O. (2012). *Mplus Version 7 User's Guide*. Los Angeles, CA: Muthén & Muthén.

Nunnally, J. C. (1978). *Psychometric Theory*, 2nd Edit. Hillsdale, NJ: McGraw-Hill.

Orlando, M., and Thissen, D. (2003). Further investigation of the performance of S-X2: an item fit index for use with dichotomous item response theory models. *Appl. Psychol. Meas.* 27, 289–298. doi: 10.1177/0146621603027004004

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *J. Educ. Stat.* 4, 207–230. doi: 10.3102/10769986004003207

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med. Care* 45, S22–S31.

Reise, S. P., and Waller, N. G. (2009). Item response theory and clinical measurement. *Annu. Rev. Clin. Psychol.* 5, 27–48. doi: 10.1146/annurev.clinpsy.032408.153553

Revicki, D. A., and Cella, D. F. (1997). Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. *Qual. Life Res.* 6, 595–600.

Rice, M. E., and Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law Hum. Behav.* 29, 615–620. doi: 10.1007/s10979-005-6832-7

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph* 17, 5–17. doi: 10.1007/s11336-012-9273-5

Schisterman, E. F., Perkins, N. J., Liu, A., and Bondell, H. (2005). Optimal cut-point and its corresponding youden index to discriminate individuals using pooled blood samples. *Epidemiology* 16, 73–81. doi: 10.1097/01.ede.0000147512.81966.ba

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Smits, N., Cuijpers, P., and van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: a simulation study. *Psychiatry Res.* 188, 147–155. doi: 10.1016/j.psychres.2010.12.001

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (1998). *Bayesian Deviance, the Effective Number of Parameters, and the Comparison of Arbitrarily Complex Models. Research Report, 98–009.* Available at: http://www.med.ic.ac.uk/divisions/60/biointro.asp (accessed February 2018).

Tan, Q., Cai, Y., Li, Q., Zhang, Y., and Tu, D. (2018). Development and validation of an item bank for depression screening in the chinese population using computer adaptive testing: a simulation study. *Front. Psychol.* 9:1225. doi: 10.3389/fpsyg.2018.01225

Thatcher, A., and Goolam, S. (2005). Development and psychometric properties of the problematic internet use questionnaire. *S. Afr. J. Psychol.* 35, 793–809. doi: 10.1177/008124630503500410

Tinsley, H. E. (1972). *An Investigation of the Rasch Simple Logistic Model for Tests of Intelligence or Attainment*.Ph.D. thesis, ProQuest Information & Learning, Michigan.

Tran, B. X., Hinh, N. D., Nguyen, L. H., Le, B. N., Nong, V. M., Thuc, V. T. M., et al. (2017a). A study on the influence of internet addiction and online interpersonal influences on health-related quality of life in young Vietnamese. *BMC Public Health* 17:138. doi: 10.1186/s12889-016-3983-z

Tran, B. X., Mai, H. T., Nguyen, L. H., Nguyen, C. T., Latkin, C. A., Zhang, M. W., et al. (2017b). Vietnamese validation of the short version of internet addiction test. *Addict. Behav. Rep.* 6, 45–50. doi: 10.1016/j.abrep.2017.07.001

van der Ark, L. A. (2007). Mokken scale analysis in R. *J. Stat. Softw.* 20, 1–19.

Wainer, H., Bradlow, E. T., and Du, Z. (2000a). *Testlet Response Theory: An Analog for the 3PL Model Useful in Testlet-Based Adaptive Testing. In Computerized Adaptive Testing: Theory and Practice*. Netherlands: Springer, 245–269.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., and Mislevy, R. J. (2000b). *Computerized Adaptive Testing: A Primer*. Abingdon: Routledge.

Walter, O. B., Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F., and Rose, M. (2007). Development and evaluation of a computer adaptive test for 'Anxiety' (Anxiety-CAT). *Qual. Life Res.* 16, 143–155. doi: 10.1007/s11136-007-9191-7

Wang, T., Hanson, B. A., and Lau, C. M. A. (1999). Reducing bias in CAT trait estimation: a comparison of approaches. *Appl. Psychol. Meas.* 23, 263–278. doi: 10.1177/01466219922031383

Wang, T., and Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *J. Educ. Meas.* 35, 109–135. doi: 10.1111/j.1745-3984.1998.tb00530.x

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika* 54, 427–450. doi: 10.1007/bf02294627

Weiss, D. J., and Kingsbury, G. (1984). Application of computerized adaptive testing to educational problems. *J. Educ. Meas.* 21, 361–375. doi: 10.1177/0146621617707509

Wickham, H. (2011). ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* 3, 180–185. doi: 10.1002/wics.147

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *J. Educ. Meas.* 30, 187–213. doi: 10.1111/j.1745-3984.1993.tb00423.x

Young, J. W. (1990). Adjusting the cumulative GPA using item response theory. *J. Educ. Meas.* 27, 175–186. doi: 10.1111/j.1745-3984.1990.tb00741.x

Young, K. (1998). *Caught in the Net*. New York, NY: John Wiley.

Zhang, M. W., Lim, R. B., Lee, C., and Ho, R. C. (2018). Prevalence of internet addiction in medical students: a meta-analysis. *Acad. Psychiatry* 42, 88–93. doi: 10.1007/s40596-017-0794-1

Zhang, M. W., Tran, B. X., Hinh, N. D., Nguyen, H. L. T., Tho, T. D., Latkin, C., et al. (2017). Internet addiction and sleep quality among Vietnamese youths. *Asian J. Psychiatry* 28, 15–20. doi: 10.1016/j.ajp.2017.03.025

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Validation of Embedded Experience Sampling (EES) for Measuring Non-cognitive Facets of Problem-Solving Competence in Scenario-Based Assessments

Andreas Rausch[1]*, *Kristina Kögler*[2] *and Jürgen Seifried*[3]

[1] Economic and Business Education – Workplace Learning, Business School, University of Mannheim, Mannheim, Germany, [2] Business Education, University of Hohenheim, Stuttgart, Germany, [3] Economic and Business Education – Professional Teaching and Learning, Business School, University of Mannheim, Mannheim, Germany

OPEN ACCESS

Edited by:
Ronny Scherer,
University of Oslo, Norway

Reviewed by:
Samy A. Azer,
King Saud University, Saudi Arabia
Christian Wandeler,
California State University, Fresno,
United States

***Correspondence:**
Andreas Rausch
rausch@uni-mannheim.de

Specialty section:
This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Psychology

Received: *07 December 2018*
Accepted: *07 May 2019*
Published: *24 May 2019*

Citation:
Rausch A, Kögler K and Seifried J
(2019) Validation of Embedded
Experience Sampling (EES)
for Measuring Non-cognitive Facets
of Problem-Solving Competence
in Scenario-Based Assessments.
Front. Psychol. 10:1200.
doi: 10.3389/fpsyg.2019.01200

To measure non-cognitive facets of competence, we developed and tested a new method that we refer to as Embedded Experience Sampling (EES). Domain-specific problem-solving competence is a multi-faceted construct that is not limited to cognitive facets such as domain knowledge or problem-solving strategies but also comprises non-cognitive facets in the sense of domain-specific emotional and motivational dispositions such as, for instance, interest and self-concept. However, in empirical studies non-cognitive facets are usually either neglected or measured by generalized self-report questionnaires that are detached from the performance assessment. To enable an integrated measurement, we developed the EES method to collect data on non-cognitive facets during scenario-based low-stakes assessments. Test-takers are requested to stop at certain times and spontaneously answer short items (EES items) regarding their actual experience of the problem situation. These EES items are embedded in an EES event that resembles typical social interactions with non-player characters. To evaluate the feasibility and validity of the method, we implemented EES in a series of three studies in the context of commercial vocational education and training (VET): A feasibility study with 77 trainees, a pilot study with 20 trainees, and the main study with 780 trainees who worked on three complex problem scenarios in a computer-based office simulation. In the present paper, we investigate how test-takers perceived the EES events, and whether social desirability biased their answers, and investigate the internal structure of the data and the relationship between EES data and data from several other sources. Interview data and survey data indicated no biases due to social desirability and no additional burden for the test-takers due to the EES events. A correlation analysis following the multitrait-multimethod approach as well as the calibration of a multidimensional model based on Item Response Theory (IRT) also supported the construct validity. Furthermore, EES data shows substantial correlations with test motivation but almost zero correlations with data from generalized retrospective self-report questionnaires on non-cognitive facets. Altogether, EES offers an alternative approach to measuring non-cognitive facets of competence under certain conditions. For instance, EES is also based on self-reporting and thus might not be suitable for high-stakes testing.

Keywords: embedded experience sampling, competence assessment, non-cognitive facets, problem solving, computer-based assessment, scenario-based assessment, business simulation

INTRODUCTION

Problem-solving competence has gained increasing attention in educational science as well as in vocational education and training (VET) and professional development. In vocational and professional contexts, problem-solving competence is important because of a general trend toward higher-order skills owing to the ongoing automatization and outsourcing of routine tasks that not only affect blue-collar work in production lines but also white-collar work (e.g., Brynjolfsson and McAfee, 2014; Frey and Osborne, 2017). Problem solving is considered to be an orchestration of cognitive, metacognitive, and non-cognitive processes in order to find an initially unknown way of bridging the gap between an actual state and a desired state (Dörner and Funke, 2017). Hence, unlike routine action, problem solving is by definition strenuous and problems usually evoke negative emotions that have to be dealt with. Altogether, problem solving is enhanced by motivation, excitement, perseverance, frustration tolerance, emotion regulation, (mild) positive affect, self-confidence, and so forth (Sembill, 1992; Frensch and Funke, 1995; Sugrue, 1995; Isen, 2008; Hannula, 2015; Schoppek and Fischer, 2015). Consequently, problem-solving competence also comprises non-cognitive dispositions which are also seen to be part of competence in general and work competence more specifically (Weinert, 2001; Rychen and Salganik, 2003; Kanfer and Ackerman, 2005). Nevertheless, the assessment of competencies is usually limited to cognitive aspects such as the reproduction or application of domain knowledge. We argue that a more holistic assessment of problem-solving competence should result in a competence profile that also comprises non-cognitive facets (Sembill et al., 2013; Rausch and Wuttke, 2016). The lack of holistic measurement approaches has led us to develop an experience sampling procedure which builds on the integration of emotional and motivational self-reports into computer-based competence assessments. It is referred to as Embedded Experience Sampling (EES) and has been created to capture the non-cognitive dimension of problem solving *in situ*. This contribution outlines the characteristics and implementation of EES and presents findings concerning its validity gained by conducting three empirical studies throughout the developmental process.

Non-cognitive Facets of Problem-Solving Competence

In his seminal report, Weinert (2001) developed a broad definition of action competence as a combination of "intellectual abilities, content-specific knowledge, cognitive skills, domain-specific strategies, routines and subroutines, motivational tendencies, volitional control systems, personal value orientations, and social behaviors" (Weinert, 2001, p. 51). He pointed out that "performance in specific situations depends on more than cognitive prerequisites" (Weinert, 1999, p. 19). Similarly, Kanfer and Ackerman (2005) consider knowledge, skills, abilities, motivation, personality, and self-concept as components of work competence. Furthermore, within research on problem solving, there is a broad consensus that besides the significance of domain-specific knowledge, problem solving is also enhanced by "... some non-cognitive factors such as self-confidence, perseverance, motivation, and enjoyment" (Frensch and Funke, 1995, p. 21). Within the framework of problem solving introduced by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), problem-solving competence comprises motivation (further divided into effort and self-efficacy) along with cognitive facets (Herl et al., 1999). Similar definitions are found in research on mathematical problem solving (Verschaffel et al., 2012; Schoenfeld, 2013). There is no universally accepted definition of the term "non-cognitive" (Duckworth and Yeager, 2015) just as there is no such definition of "cognition" (Neisser, 1967). Any attempt to distinguish cognitive from non-cognitive constructs remains artificial, but facilitates the understanding and analysis of their interdependence (Weinert, 1999).

When solely focusing on the assessment of cognitive facets of competence, it is implicitly assumed that test-takers invest maximum effort to perform as well as possible. Test performance is interpreted as maximum performance in the sense of Cronbach (1960) and thus varying test motivation threatens the validity of the assessment. It is well-known that in testing for intelligence and in international large-scale studies, test motivation exerts an influence on achievement (Butler and Adams, 2007; Duckworth et al., 2011). Eklöf (2010) points out that an achievement test score is a function of "skill and will." Correspondingly, including non-cognitive facets in the definition and modeling of competence moves the construct to be measured from "can do" to "will do" (Kanfer and Ackerman, 2005; Cortina and Luchman, 2012); or, respectively, from maximum performance to typical performance in the sense of Cronbach (1960). Consequently, emotions and motivation no longer represent construct-irrelevant variance, but are a manifest result of latent non-cognitive facets of competence which has to be considered in the measurement. Regarding convergent validity, data of non-cognitive facets of competence should be correlated with measures of test motivation.

Based on a literature review, we developed a competence model that distinguishes knowledge application, action regulation, self-concept, and interest as components of

TABLE 1 | Model of domain-specific problem-solving competence (Rausch and Wuttke, 2016, p. 177).

Four components of competence	Thirteen facets of competence			
(A) Knowledge application(cognition)	Identifying needs for action and information gaps	Processing information	Coming to well-founded decisions	Communicating decisions appropriately
(B) Action regulation(metacognition)	Planned (well-structured) action	Persistence (focused action)		Retrospective action control
(C) Self-concept(expectancies)	Situational confidence in one's competence	Ambiguity/uncertainty tolerance		Situational confidence in one's solution
(D) Interests(valences)	Personal interest in the problem context/content	Maintaining positive and active emotional states		Interest in the progress of/in learning from the problem

domain-specific problem-solving competence (**Table 1**). We further defined several facets within each of the components. These facets are arranged alongside an ideal problem-solving process and are intended to guide the measurement of problem-solving competence (Rausch and Wuttke, 2016).

The non-cognitive components (self-concept and interest) mirror the expectancy-value theory of achievement motivation (Wigfield and Eccles, 2000) and the control and value appraisals of achievement motivation (Pekrun, 2006), respectively. Confidence in one's own competence when confronted with a domain-specific problem, tolerating ambiguity and uncertainty, and having confidence in one's own solutions concerning domain-specific problems are defined as facets of a domain-specific self-concept. Being interested in the context of a domain-specific problem, maintaining positive and active emotional states while working on a domain-specific problem, and being interested in the progress of and learning from these problems are defined as facets of domain-specific interest.

Modeling and Measuring Non-cognitive Facets of Competence

Based on a multidimensional understanding of competence, a crucial question is how non-cognitive facets are measured. Two basic options in dealing with the multidimensionality of the construct can be distinguished (Sembill et al., 2013).

Multifaceted Competence Model With Fragmented Measurement

Following this very common approach, non-cognitive facets are part of a multifaceted construct of competence but are measured separately, usually by administering retrospective self-report questionnaires. Those self-reports remain detached from the actual performance. In general, self-reports are considered face-valid (Debus, 2000) but there is plenty of research that stresses several threats and biases regarding the validity of decontextualized retrospective self-reports on emotion and motivation (van Reekum and Scherer, 1997; Robinson and Clore, 2002; Novak and Johnson, 2012; Schwarz, 2012). Furthermore, in their investigation of the empirical relation between intelligence and problem solving, Wittmann and Süß (1999) point to the "Brunswik asymmetry" named after Brunswik (1956) in order to explain the poor prediction of problem solving via intelligence. This poor relation is due to an asymmetry in the content and breadth of the predictor (intelligence) and the criterion (problem solving), because the former is a very broad construct, while the latter is derived from a contextualized performance task. The same argument holds true for the relation of problem solving and non-cognitive facets if non-cognitive facets are measured through general self-report questionnaires which are detached from problem solving (Rausch et al., 2016; Rausch, 2017). This approach may lead to an underestimation of the importance of non-cognitive competence facets (Dermitzaki et al., 2009; Sembill et al., 2013).

Multifaceted Competence Model With an Integrated Measurement

Following an integrated approach, the measurement of non-cognitive facets is integrated into the performance assessment. Regarding the differentiation of state and trait, recurrent situational emotional states are interpreted as the dispositional core of a trait emotion (Diener and Lucas, 2000). Just as the assessment of cognitive facets of competence is based on the repeated measurement of manifest performance, the suggested *in situ* assessment of non-cognitive facets is based on the repeated measurement of emotional states in the context of different problem scenarios. A multitrait-multimethod approach (MTMM; Campbell and Fiske, 1959) can be applied to investigate the internal or construct validity of such an approach. The multiple problem scenarios constitute different methods and the various non-cognitive facets (see **Table 1**) constitute different traits. According to MTMM (Podsakoff et al., 2003), higher correlations between the same traits across different scenarios (monotrait-heteromethod) than between different traits within one scenario (heterotrait-monomethod) indicate internal or construct validity.

Embedded Experience Sampling to Measure Non-cognitive Facets of Competence

Our empirical approach to measuring non-cognitive facets of competence is inspired by the Experience Sampling Method (ESM) which was introduced by Csikszentmihalyi and Larson (1987, p. 526) as "an attempt to provide a valid instrument to describe variations in self-reports of mental processes.". In ESM, participants are repeatedly requested to report their emotional states over a period of time. Different types of ESM have been established (Scollon et al., 2003, p. 7ff.): Signal-contingent sampling requires participants to complete self-reports when prompted by a randomly-timed signal (e.g., twice a day). Event-contingent sampling requires participants to complete self-reports whenever a predefined event occurs (e.g., in case of problems). Interval-contingent sampling uses constant time-intervals. The Continuous State Sampling Method (CSSM) is a special case of such time-sampling ESM with very short intervals of only 5–10 min. CSSM has been developed and applied in the context of classroom research (Sembill et al., 2008; Conrad and Schumann, 2017; Kärner et al., 2017; Kögler and Göllner, 2018). CSSM is also used for validating our own approach.

Our development of Embedded Experience Sampling (EES) builds on traditional ESM. In order to measure the non-cognitive facets in computer-based tests on problem-solving competence, EES aims at collecting self-report data on non-cognitive facets *in situ* and furthermore integrates these self-reports into the storyline of authentic problem scenarios. Test-takers are briefly interrupted during the test and requested to answer short questions (EES items) regarding their momentary experience. These EES items are embedded into the test situation in authentic EES events that resemble ordinary social interaction at the workplace (e.g., a colleague asks how one is doing). Closed-ended questions were used in order to spare the test-takers the time they would need to write down their answers. Furthermore, they improve the comparability of the answers and facilitate the implementation of EES in large-scale assessments regarding psychometric scaling. EES items focus on difficult to monitor non-cognitive competence facets such as interest, attitudes, commitment, and self-concept.

A similar approach was applied in PISA 2006 as an "embedded science interest assessment". Directly after working on selected test items regarding science competence, the participants were requested to rate their situational interest in the prior item context. The data were calibrated in Item Response Theory (IRT) models to assess trait interest (Drechsel et al., 2011). However, few such approaches are so far known to the authors. Furthermore, the EES approach differs from the PISA approach because in PISA the items were not embedded into the "storyline" of the assessment. A further example for integrating experience sampling into a complex assessment is the "affect self-report device" applied to the game-based learning environment "Crystal Island." During their interaction with the learning environment, test-takers received an in-game prompt asking them to report on their cognitive and emotional states. These status updates were described as part of an in-game social network (Sabourin and Lester, 2014). The "affect self-report device" is embedded in the sense of EES, but it was not designed to measure non-cognitive traits as part of a competence assessment.

Any sampling of self-reported experiences *in situ* faces limitations: for instance, social desirability may affect individuals' responses and possibly lead to a bias in the psychometric data in terms of construct-irrelevant variance (Messick, 1994). In this context, the criteria of cognitive validity (Pellegrino et al., 2016) or construct validity (Messick, 1994), respectively, require that participants do not consciously deliberate about whether a particular answer would be more socially desirable but only answer according to their actual situational experience. Following the argument of Reis (2012), measuring non-cognitive facets within the problem-solving process promotes ecological validity, given that the problem scenarios and the EES events are representative of daily work. Furthermore, biases due to social desirability might decrease in EES compared to retrospective self-reports, due to the concurrent cognitive load and time pressure during the problem-solving process (Stodel, 2015). However, the repeated sampling of subjective states may also cause reactivity and reactance, for better or worse, because on the one hand it constitutes a disruption and on the other hand it may also trigger reflection (Csikszentmihalyi and Larson, 1987; Scollon et al., 2003; Novak and Johnson, 2012).

Research Questions and Hypotheses

We implemented EES into test situations in three field studies and collected EES data to investigate

- How test-takers perceived the EES events (RQ1),
- Whether social desirability biased their answers (RQ2),
- The internal structure of the data (RQ3) and
- The relationship between EES data and (a) CSSM data, (b) test motivation, and (c) generalized retrospective self-reports (RQ4).

Table 2 gives an overview of the research questions and corresponding hypotheses of the field studies.

The studies were part of the research project 'modellng and measuring domain-specific problem-solving competence of industrial clerks' (DomPL-IK), which was funded by the German Federal Ministry of Education and Research (Grant No. 01DB081119–01DB1123). The apprenticeship program to become an industrial clerk is the fifth most frequent of nearly 330 state-recognized apprenticeship programs in the well-respected German dual system of vocational education and training (VET). Apprenticeship programs usually require 3 years to complete and are characterized by a combination of workplace learning in the training company and classroom-based learning in state-run vocational schools. Certified industrial clerks usually work in back-office departments of industrial or service companies. A general description of the research project and selected results have been published in Rausch et al. (2016).

In the present article, we focus on the development and validation of the EES approach by analyzing EES data from two pilot studies and the main study. In a first feasibility study, we investigated how participants perceived the EES events, whether social desirability played a role, whether the EES data met the requirements of the MTMM approach, and how EES data were correlated to retrospective measures of interest and self-concept. The aims of the second pilot study was to test the computer-based office simulation that, for the first time, also included a computer-based implementation of EES events. Additional data were collected to investigate the subjective experience of the EES, social desirability in EES responses, and the relation to CSSM data and test motivation. Finally, the computer-based assessment of domain-specific problem-solving competence was implemented in a large-scale study with almost 800 participants in vocational schools in six federal German states. The resulting EES data were calibrated in a psychometric model based on Item Response Theory (IRT). Parts of this final step of the test development are published in

TABLE 2 | Overview of the studies, the research questions, and the hypotheses.

Research questions	Study 1	Study 2	Study 3
RQ1 (ecological validity): test-takers' perception of EES	H1a: Participants in low-stake tests do not experience EES events as an additional and unrealistic burden (interview study).	H1b: Participants in low-stake tests do not experience EES events as an additional and unrealistic burden (survey data).	
RQ2 (construct validity): social desirability	H2a: Participants in low-stake tests answer EES items without deliberating about desirable answers (interview study).	H2b: EES data and scores from the Balanced Inventory of Desirable Responding (BIDR) show zero to small correlations.	
RQ3 (construct validity): structure of the data	H3a: EES data meets the requirements of the multitrait-multimethod (MTMM) approach.		H3b: EES data meets the requirements of a multidimensional model based on Item Response Theory (IRT).
RQ4: relations between EES and			
(a) . . . CSSM data (convergent validity)		H4a: EES data of situational interest and CSSM data of situational interest show medium to large correlations.	
(b) . . . test motivation (convergent validity)		H4b: EES data of situational interest and test motivation show medium to large correlations.	
(c) . . . generalized retrospective self-reports (divergent validity)	H4c: EES-based scores and retrospective measures of vocational interest and self-concept show small correlations.		H4c (Replication): EES-based scores and retrospective measures of vocational interest and self-concept show small correlations.

Rausch et al. (2016). The studies within the research project have been approved by the responsible ministries of education and the responsible commissions of data protection of the respective German Federal States as well as by the Ethics Committee of the Otto-Friedrich-University of Bamberg (Otto-Friedrich-University Bamberg, Bamberg, Germany).

STUDY 1: FEASIBILITY STUDY OF IMPLEMENTING EES EVENTS INTO AUTHENTIC PROBLEM SCENARIOS

Materials and Methods

Participants

The feasibility of implementing EES in the assessment of domain-specific problem-solving competence was investigated in a pilot study with $N = 77$ students in vocational education and training (VET) of two vocational business schools in Germany. All participants were enrolled in a 3-year apprenticeship program to become industrial clerks and were nearing the end of their 2nd year of the apprenticeship. The sample included 28 male and 49 female participants who showed a typical age distribution ($M = 21.8$; $SD = 1.56$; min = 18; max = 26). Participation was voluntary and all participants provided written informed consent.

Procedure

Data were collected in computer-equipped classrooms. At the beginning of the data collection sessions the researchers introduced themselves, the project, and the agenda. First, the participants completed several self-report questionnaires including scales on vocational interest and work-related self-efficacy. Next, they worked on three authentic, computer-based business problems including the completion of several EES items (for further information see Rausch, 2017). The session ended with group discussions or individual interviews about the problem scenarios and the experience of EES.

The three computer-based problem scenarios required a cost deviation analysis (30 min), a supplier selection (40 min), and a make-or-buy decision (50 min). Each scenario started with an email from a supervisor which included a problem and a variety of documents of varying relevance, transparency, and credibility. All scenarios required participants to go through multiple processes of information seeking, processing, and interpreting. To complete a scenario, the participants had to reply to the initial email with a well-founded proposed solution. The test environment provided "open book" conditions meaning that participants could look up technical terms, formulae, legal regulations etc. in a large reference work. However, they were not allowed to consult any other sources such as the internet. The participants used Microsoft Excel® to work on several spreadsheet files and Microsoft Word® documents to write their email reply and make notes. The problem environment was open in the sense that there was no further structure provided during the given time frame for each problem scenario. Editable documents were analyzed for each participant to assess the cognitive facets of problem-solving competence (see **Table 1**). For further information on the analysis of the cognitive facets see Rausch et al. (2016), Rausch (2017), and Seifried et al. (unpublished).

Measures

Embedded experience sampling (EES)

In this feasibility study, four EES events were implemented into each of the above problem scenarios. **Table 3** lists the EES

TABLE 3 | Overview of EES events, competence facets, and EES items in Study 1.

EES event (point of time)	Competence facet (see Table 1)	EES items (translated from German and condensed)
Short email response after the reception of the task (after 3 min)	Situational confidence in one's competence (C1)	C1_1: Sender of the task requests a first quick estimation. Answer from 1 = '*I do not know what to do here yet*' to 4 = '*I know exactly what to do here.*'
Phone call from the sender of the task (after 10 min; in scenario 3 after 20 min)	Situational confidence in one's competence (C1)	C1_2: Sender of the tasks requests a further estimation. Answer from 1 = '*I am afraid that I will not be able to cope with it, but I will do my best*' to 4 = '*I can definitely cope with it and I will do my best.*'
Short visit by a colleague (after 20 min; in scenario 3 after 35 min)	Maintaining positive and active emotional states (D2)	Friend enters the office asks how one is doing. D2_1: from 1 = '*not at all*' to 4 = '*very nervous/worried.*' (–) D2_2: from 1 = '*not at all*' to 4 = '*very motivated/interested.*' D2_3: from 1 = '*not at all*' to 4 = '*very irritated/annoyed.*' (–) D2_4: from 1 = '*not at all*' to 4 = '*very confident/optimistic.*'
Short request from the sender of the task after the reception of the solution (after submission or after 30 min in scenario 1, after 40 min in scenario 2 and after 50 min in scenario 3, respectively)	Situational confidence in one's solution (C3)	C3: Sender of the task asks how confident the apprentice is about her/his solution and whether the solution has to be checked before its implementation. Answer from 1 = '*Unfortunately*, *I did not arrive at a solution at all*' over 2 = '*I am afraid you should check everything in detail because I assume I made some mistakes*' to 5 = '*I think I found a proper solution that you do not have to check in detail again.*'
	Interest in the progress of/in learning from the problem (D3)	Participants can add none, several or all of the following statements to his/her e-mail answer. '*I would be very happy if you could . . .*' D3_1: . . . *inform me about the final decision that you made.*' 'D3_2: . . . *give me feedback in case of any errors I made*'. D3_3: . . . *explain the correct procedure to me*'. D3_4: . . . *assign similar cases to me in the near future*'.

EES events and EES items were the same for all of the three problem scenarios; (–) indicate inverse items. Abbreviations behind competence facets refer to competence model in ***Table 1****; facets D1 and C2 have not been measured yet in this first study.*

events, the related competence facets, and the EES items that were used. In this first application of the method, no events and items had been designed for the competence facets C2 "ambiguity/uncertainty tolerance" and D1 "personal interest in the problem context/content".

In this early stage of the project, EES events were paper-based and came in separate envelopes that were numbered consecutively and placed on each participant's desk (see **Appendix Figure A1** for an example). Female and male participants were provided a gender-specific version of the EES events. At predefined times during the test, participants were asked to open a particular envelope, to immediately complete the items, and to put the paper sheet back into the envelope. Altogether, 1,845 such envelopes were prepared for this study. Apparently, test efficiency was questionable in this paper-based implementation of EES.

The data of the two EES items concerning the competence facet "confidence in one's competence" (C1) were condensed into one scale for each scenario. The internal consistencies were not satisfactory (0.57 < Cronbach's alpha < 0.59). "Situational confidence in one's solution" (C2) was measured with a single item (see **Table 3**). The data of the four EES items on the competence facet "positive and active emotional state" (D2) were condensed into one scale for each scenario. Inverse items were re-coded and a mean score was calculated for each scenario. Again, the internal consistencies were not satisfactory (0.56 < Cronbach's alpha < 0.61). The four dichotomous EES items on the competence facet "interest in the progress of the problem" (D3) were condensed into one scale for each scenario by sum score. Thus, the scores for each non-cognitive facet ranged from 1 to 4.

Generalized self-reports of work-related self-efficacy and work-related interest

We administered a scale designed to measure work-related self-efficacy (Abele et al., 2000). The scale consisted of six statements that were rated on a five-point Likert scale ranging from 1 = disagree to 5 = agree (e.g., "I do not worry about work-related challenges because I can always trust my abilities."). The internal consistency of the scale was satisfactory (Cronbach's alpha = 0.73). An adapted and shortened version of a scale originally developed to measure dispositional interests in students (Schiefele et al., 1993) was administered. The scale consisted of six statements rated on a four-point Likert scale ranging from 1 = disagree to 4 = agree. The items assessed general interest in the current apprenticeship program (e.g., "I am sure that I have chosen an apprenticeship program which reflects my personal interests."). The internal consistency of the scale was satisfactory (Cronbach's alpha = 0.76).

Subjective experience of EES

To investigate how the participants experienced the EES, two group discussions in class (with approximately 20 participants each) and 11 individual interviews were conducted. Participants were asked how they experienced the procedure (the additional questions that came in the envelopes). They were asked whether they had deliberated about alternative responses and whether answering these questions had caused additional stress during their work on the problem situations.

Data Analysis

Following a multitrait-multimethod (MTMM) approach, the various facets of competence are multiple traits and the three scenarios are multiple methods. Although the variables were not normally distributed (Shapiro–Wilk tests), parametric Pearson correlations were calculated since this method is considered robust (Norman, 2010). In correlation tables, indications of significance are omitted in favor of legibility. Following Cohen (1988), correlation coefficients of $0.10 < r < 0.30$ indicate small effects, $0.30 < r < 0.50$ indicate medium effects, and $r > 0.50$ indicate large effects. The interview data were categorized with regard to social desirability and the additional burden of answering the EES items while working on the problem scenarios. The data was analyzed using IBM SPSS 24.

Results

Descriptive Statistics

The mean values for the EES variables range between 1.71 and 3.19 on a four-step scale (see **Appendix Table A1**). The variable D2 (maintaining positive and active emotional states) shows high values, consistently above the value of 2.3, while variable D3 (interest in the progress of/in learning from the problem) shows much lower values. Here, the mean values only reach a value above 2.0 in scenario 2. Finally, the decrease of the mean values over time for variable C3 (situational confidence in one's solution) is noteworthy. The mean value drops from 2.97 in scenario 2 to 1.71 in scenario 3. This finding is in line with the difficulty of the scenarios (determined by the solution rates)—scenario 2 was evaluated as the easiest one while scenario 3 showed the lowest solution rate, as expected with regard to the complexity of the scenario.

Test-Takers' Perception and Social Desirability (RQ1, RQ2)

To investigate participants' subjective experience of the EES, individual interviews and group discussions were conducted. In both group discussions the participants reacted positively to the way in which social interaction was implemented via the paper-based questionnaires and stated that such interruptions were quite realistic. Two of the 11 individually interviewed participants made similar statements when asked how they experienced these short questionnaires and added that it was an entertaining addition to the test scenarios. None of the participants reported adverse experiences. In one group discussion, a participant cautiously indicated that one could have thought about how some of the responses would appear to others. All of the 11 individually interviewed participants indicated that they answered spontaneously according to their actual experience and did not deliberate about "good answers". Only one out of 11 participants stated that answering the EES items caused an additional burden. Altogether, the participants' responses gave no reasons to assume biases from social desirability or any additional burden and thus they support H1a and H2a (see **Table 2**).

Multitrait-Multimethod Analyses (RQ3)

In the next step, we analyzed the structure of the data by applying a multitrait-multimethod approach. High heterotrait-monomethod correlations between different non-cognitive competence facets (traits) within a scenario (method) argue for situational influences of the scenario, while high monotrait-heteromethod correlations between the same competence facets (traits) measured in different scenarios (method) argue for trait influences. **Table 4** shows the results of the MTMM analysis.

The mean correlation of all 18 heterotrait-monomethod combinations is $r = 0.28$ while the mean correlation of all 12 monotrait-heteromethod combinations is $r = 0.33$, which is consistent with the MTMM assumption. Heterotrait-monomethod correlations different from zero are plausible because the theoretical constructs are not assumed to be fully independent of each other. The monotrait-heteromethod correlations are higher which supports the assumption of internal validity and thus supports H3a (see **Table 2**). However, they are not much higher than the heterotrait-monomethod correlations. Internal consistency across all three scenarios and across both EES variables of self-concept was CA = 0.66 (6 variables); the respective internal consistency across all three scenarios and across both EES variables of interest was CA = 0.71 (6 variables).

Relations Between EES and Generalized Retrospective Self-Reports (RQ4)

Finally, by calculating mean scores across the EES variables, we received two EES-based scales, one for self-concept and one for interest. The correlations between EES-based scales and scales from generalized self-reports of work-related self-efficacy and vocational interest were close to zero and not significant ($r = 0.05$, $p = 0.66$ for self-concept; $r = 0.04$, $p = 0.69$ for interest). We hypothesized small correlations (H4c) even though the theoretical constructs are quite similar.

STUDY 2: VALIDATION STUDY OF RESPONSES TO COMPUTER-BASED EES EVENTS

Materials and Methods

Participants

To test the computer-based implementation of EES events and the subjective experience of the EES, 21 VET students participated voluntarily in this pilot study and provided written informed consent. Eight participants were male and 13 were female; the participants were 20.3 years old on average (*SD* = 1.93; min = 18; max = 24).

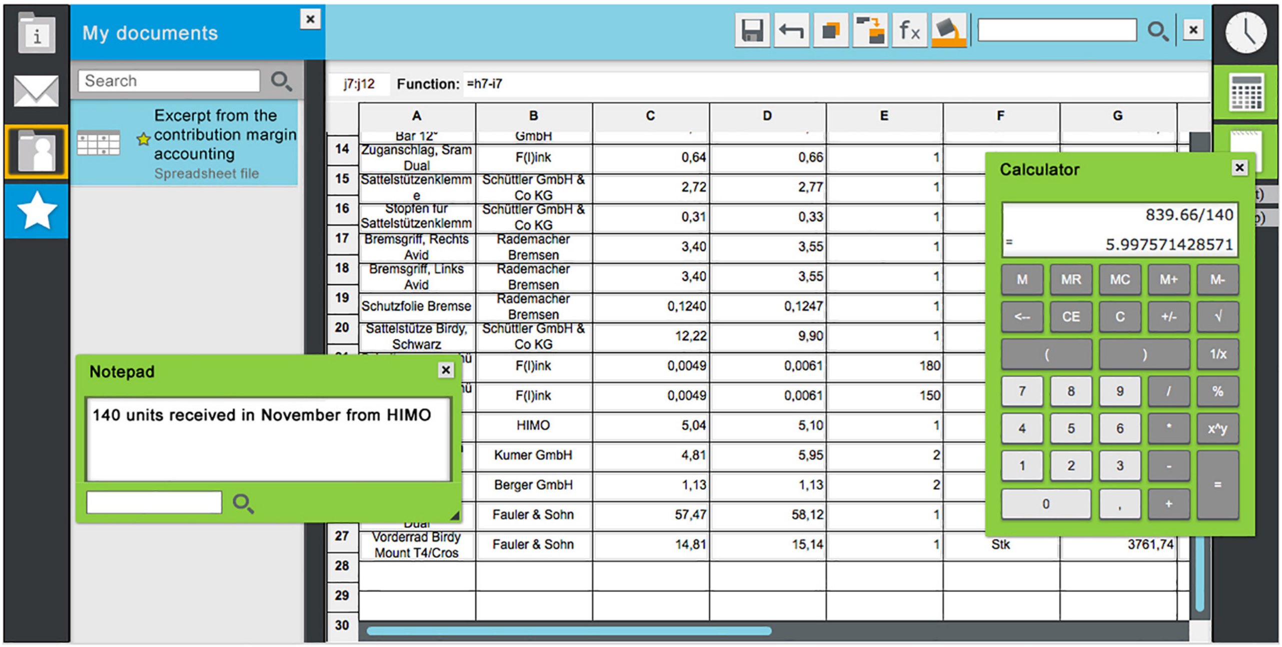
Procedure

Data were collected in a computer-equipped classroom. At the beginning of the sessions the researchers introduced themselves, the project, and the agenda. The participants worked on one authentic, computer-based problem scenario including the completion of several EES items. In contrast to the feasibility study, the scenario in this pilot study was presented and completed in an integrated custom-built office simulation that comprised typical features of an office workplace, such as an email client, a spreadsheet application, a folder structure, a file viewer, a notepad, a calculator and so forth. **Figure 1** shows a screenshot of the office simulation.

TABLE 4 | Correlations between EES items within and across three problem scenarios in Study 1.

	Scenario 1				Scenario 2				Scenario 3			
	C1	**C3**	**D2**	**D3**	**C1**	**C3**	**D2**	**D3**	**C1**	**C3**	**D2**	**D3**
Scenario 1												
C1	1.00											
C3	0.32	1.00										
D2	0.34	0.43	1.00									
D3	–0.06	0.16	–0.08	1.00								
Scenario 2												
C1	0.29				1.00							
C3		0.24			0.40	1.00						
D2			0.20		0.42	0.40	1.00					
D3				0.71	0.15	0.07	0.10	1.00				
Scenario 3												
C1	0.05				0.46				1.00			
C3		0.08				0.23			0.64	1.00		
D2			0.29				0.24		0.45	0.54	1.00	
D3				0.58				0.64	0.31	0.22	0.21	1.00

C1, situational confidence in one's competence; C3, situational confidence in one's solution; D2, maintaining positive and active emotional states; D3, interest in the progress of/in learning from the problem; abbreviations of the competence facets refer to the competence model in ***Table 1****; facets D1 and C2 have not been measured yet in this first study.*



j7:j12 **Function:** =h7-i7

	A	B	C	D	E	F	G
	Bar 12"	GmbH					
14	Zuganschlag, Sram Dual	F(l)ink	0,64	0,66	1		
15	Sattelstützenklemme	Schüttler GmbH & Co KG	2,72	2,77	1		
16	Stopfen für Sattelstützenklemm	Schüttler GmbH & Co KG	0,31	0,33	1		
17	Bremsgriff, Rechts Avid	Rademacher Bremsen	3,40	3,55	1		
18	Bremsgriff, Links Avid	Rademacher Bremsen	3,40	3,55	1		
19	Schutzfolie Bremse	Rademacher Bremsen	0,1240	0,1247	1		
20	Sattelstütze Birdy, Schwarz	Schüttler GmbH & Co KG	12,22	9,90	1		
	nü	F(l)ink	0,0049	0,0061	180		
	nü	F(l)ink	0,0049	0,0061	150		
		HIMO	5,04	5,10	1		
		Kumer GmbH	4,81	5,95	2		
		Berger GmbH	1,13	1,13	2		
	Dual	Fauler & Sohn	57,47	58,12	1		
27	Vorderrad Birdy Mount T4/Cros	Fauler & Sohn	14,81	15,14	1	Stk	3761,74
28							
29							
30							

140 units received in November from HIMO

839.66/140

= 5.997571428571

FIGURE 1 | Screenshot of the office simulation (translated from German; Rausch et al., 2016, p. 8).

In addition to EES, data were also collected via the "Continuous State Sampling Method" (CSSM) and via a short questionnaire on test motivation and one's experience with the EES events directly after the problem scenario. Furthermore, the participants completed a longer questionnaire that included biographic information as well as several standardized scales, one of which was applied to measure a disposition toward socially desirable responding.

Measures

Embedded experience sampling method (EES)

In this pilot study of the technological implementation, four EES events were defined. However, due to a technical malfunction the fourth EES event was not presented to the participants. **Table 5** lists the remaining three EES events, the related competence facets, and the EES items that were applied.

TABLE 5 | Overview of EES events, competence facets, and EES items in Study 2.

EES event (point of time)	Competence facet (see Table 1)	EES items (translated from German and condensed)
EES event 1: short email response after the reception of the task (after 3 min)	Situational confidence in one's competence (C1)	C1_1: Sender of the task requests a first quick estimation. Answer from 1 = '*I do not know what to do here yet*' to 4 = '*I know exactly what to do here*'.
	Personal interest in the problem context/content (D1)	D1: Sender of the task asks whether tasks like this are interesting to the apprentice. Answer from 1 = '*Tasks like this are not interesting to me*' to 4 = '*Tasks like this are very interesting to me*'
EES event 2: phone call from the sender of the task (after 10 min)	Situational confidence in one's competence (C1)	C1_2: Sender of the tasks requests a further estimation. Answer from 1 = '*I am afraid that I will not be able to cope with the task, but I will do my best*' to 4 = '*I can definitely cope with the task and I will do my best*'.
	Ambiguity/uncertainty tolerance (C2)	C2: Sender of the task asks whether the apprentice likes to work on comprehensive tasks like this. Answer from 1 = '*I do not like to work on such comprehensive tasks*' to 4 = '*I very much like to work on such comprehensive tasks*'.
EES event 3: short visit by a colleague (after 20 min)	Maintaining positive and active emotional states (D2)	Friend enters the office asks how one is doing. D2_1: from 1 = '*not at all*' to 4 = '*very nervous*'. (–) D2_2: from 1 = '*not at all*' to 4 = '*very curious*'. D2_3: from 1 = '*not at all*' to 4 = '*very irritated*'. (–) D2_4: from 1 = '*not at all*' to 4 = '*very confident*'.

In this validation study, the EES events were also presented within the office simulation for the first time. **Figure 2** shows the EES event "phone call."

Embedded experience sampling data were condensed in the same way as in the feasibility study (Study 1), resulting in four EES variables for the competence facets shown in **Table 1**, C1 (confidence in one's competence), C2 (uncertainty tolerance), D1 (interest in the problem content), and D2 (positive emotional states).

Continuous state sampling method (CSSM)

Continuous state sampling method data was collected during the problem scenario via mobile devices (PalmOne Tungsten®). In 5-min intervals, the participants were requested to rate three statements on a scale from 0 to 100. The items were: (1) *Right now, this is very interesting*. (2) *Right now, I am making great efforts*. (3) *Right now, I am making great progress*. Participants were carefully instructed that this data collection was not part of the assessment and that they were expected to answer honestly according to their actual experience, while no such announcement was made for the EES events. To become familiarized with the method, the first point of measurement was before the problem scenario and was not included in the analysis. Six measurement points followed during the problem scenario at minute 4′, 9′, 14′, 19′, 24′, and 29′. Scales were calculated from the six items of each statement. Internal consistencies (Cronbach's alpha) were C.A. = 0.70 for "interesting," C.A. = 0.78 for "effort" and C.A. = 0.67 for "progress."

Social desirability

Two measures were applied to investigate whether social desirability played a role in answering the EES items. First, we administered seven items from the scale "impression management" from the "Balanced Inventory of Desirable Responding (BIDR)" (Paulhus, 1994) in a German version by Musch et al. (2002). Paulhus (1994) defined and measured "impression management" as the purposeful deception of looking good to someone. Participants were to rate statements that referred to misconduct that one is usually not willing to admit to such as, for instance, "I sometimes tell lies if I have to" (inverse item) or "I never take things that do not belong to me." Responses were given on a four-point Likert-scale. The internal consistency (Cronbach's alpha) was C.A. = 0.71. Second, immediately after the completion of the scenario, the participants completed a short questionnaire. One question aimed at "impression management" during EES responses. Participants had to rate the statement "Concerning the interposed questions, I thought hard about which answer would make me look good" on a five-point scale from 1 = *strongly disagree* to 5 = *strongly agree*.

Experience of EES

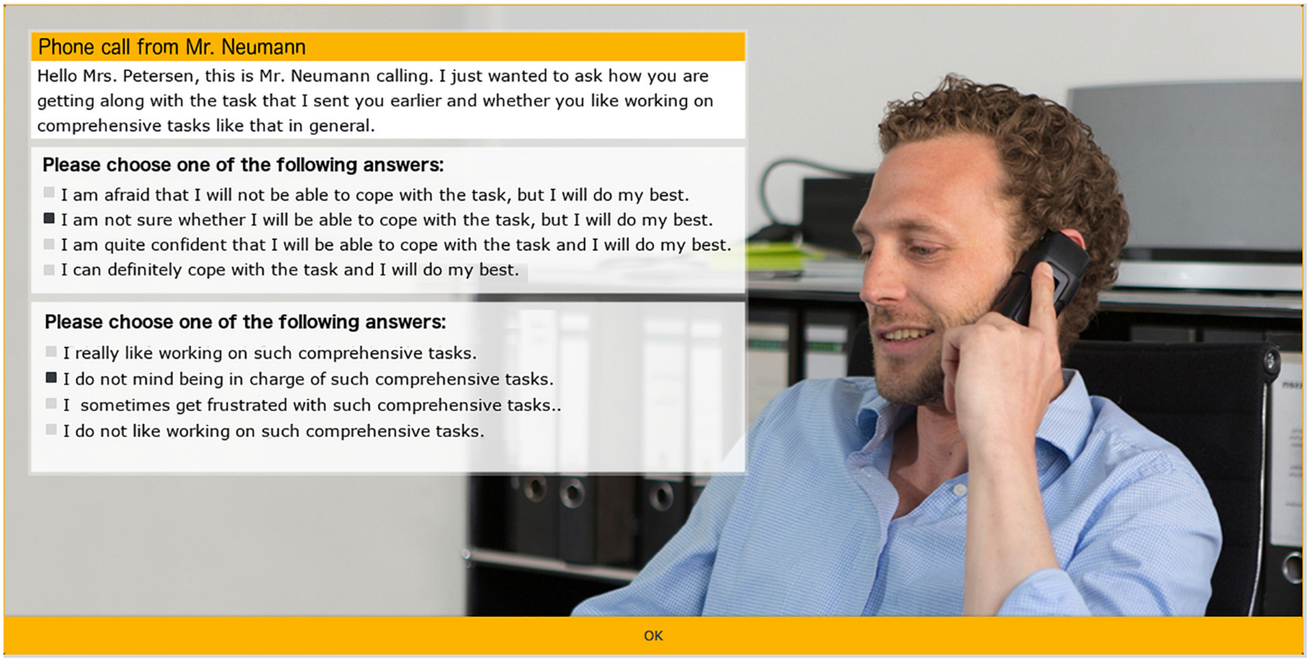
In the same short questionnaire directly after the problem scenario two additional questions were aimed at assessing the *authenticity* of the EES events ("The interposed questions [phone call, visit to my office etc.] are very realistic") and the *additional burden* due to the EES events ("I would have arrived at a better solution without these interposed questions [phone call, visit to my office etc.]").

Test motivation

We administered an adapted version of the Effort Thermometer which Kunter et al. (2002) originally developed for and applied in the Programme for International Student Assessment (PISA). The participants were requested to indicate the effort that they had invested in the previous problem scenario on a 10-point scale compared to the maximum effort they would have invested in a test situation of very high personal relevance. The Effort Thermometer was administered directly after the problem scenario.

Data Analysis

For correlation analysis, Kendall's tau-b correlations were calculated because the data were not normally distributed and the sample size was small. The data was analyzed using IBM SPSS 24.



Phone call from Mr. Neumann

Hello Mrs. Petersen, this is Mr. Neumann calling. I just wanted to ask how you are getting along with the task that I sent you earlier and whether you like working on comprehensive tasks like that in general.

Please choose one of the following answers:

- □ I am afraid that I will not be able to cope with the task, but I will do my best.
- ■ I am not sure whether I will be able to cope with the task, but I will do my best.
- □ I am quite confident that I will be able to cope with the task and I will do my best.
- □ I can definitely cope with the task and I will do my best.

Please choose one of the following answers:

- □ I really like working on such comprehensive tasks.
- ■ I do not mind being in charge of such comprehensive tasks.
- □ I sometimes get frustrated with such comprehensive tasks..
- □ I do not like working on such comprehensive tasks.

FIGURE 2 | Computer-based EES event "phone call" with two EES items (translated from German; written informed consent was obtained for the publication of this image from the individual featured).

Results

Descriptive Statistics and Social Desirability (RQ1 and RQ2)

On average, the participants experienced the problem scenario as not being very interesting (see EES variable D1 and CSSM scale "interesting"). They invested medium effort according to the CSSM scale "effort" and showed a correspondingly medium test motivation as measured by the Effort Thermometer. With regard to the EES events, the participants did not report that they tried to "look good" when answering the EES items. On average, they experienced the EES events as being quite authentic and hardly as an additional burden (see descriptive statistics in **Appendix Table A2**). Altogether, the data support H1b and H2b (see **Table 2**). The average CSSM ratings of "interesting," "progressing," and "effort" did not vary very much during the course of the problem scenario. The curve for "effort" resembles an inverted U-shape while ratings of "interesting" and "progressing" increased toward the end of the 30-min problem scenario (see **Appendix Figure A2**).

Relations Between EES Data and CSSM Data (RQ4)

Table 6 shows the correlations of selected EES items and corresponding CSSM items.

Table 6 shows that there are substantial correlations between the Embedded Experience Sampling (EES) and the Continuous State Sampling (CSSM) of situational interest (supporting H4a) while there are smaller correlations between EES data and CSSM data of confidence in one's competence and subjectively perceived progress, respectively.

Relations Between EES Data and Impression Management and Test Motivation (RQ1 and RQ4)

An analysis was made of how far EES data are influenced by social desirability or impression management and how it relates to test motivation. **Table 7** shows the results of the respective correlation analysis.

As shown in **Table 7**, there are almost zero correlations between dispositional impression management and the EES variables. Furthermore, there are only small correlations between the EES variables and situational impression management (i.e., having ". . . thought hard about which answer would make me look good"). There are medium to large correlations between some EES variables and test motivation, which is in line with our theoretical argument. Altogether, the data support H2b and H4b (see **Table 2**).

TABLE 6 | Correlations of selected EES items and corresponding CSSM items in Study 2.

	Correlations with CSSM 'Progress'						
	MP1	**MP2**	**MP3**	**MP4**	**MP5**	**MP6**	**Scale**
EES C1 (confidence)	0.48	0.17	0.25	0.25	–0.09	0.28	0.21
	Correlations with CSSM 'Interesting'						
	MP1	**MP2**	**MP3**	**MP4**	**MP5**	**MP6**	**Scale**
EES D1 (interest)	0.31	0.65	0.32	0.20	0.24	0.44	0.42

18 < n < 21. Kendal's tau-b correlations; MP, measurement point.

TABLE 7 | Correlations between EES items, impression management, and test motivation.

	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*
(1) EES C1 (confidence in one's competence)	1						
(2) EES C2 (uncertainty tolerance)	0.25	1					
(3) EES D1 (interest)	0.38	0.65	1				
(4) EES D2 (positive emotional states)	0.38	0.08	0.01	1			
(5) Dispositional impression management (BIDR)	−0.06	−0.02	0.05	−0.09	1		
(6) Situational impression management	0.17	0.28	0.22	0.29	0.01	1	
(7) Test motivation (effort thermometer)	0.29	0.57	0.62	0.30	0.17	0.26	1

17 < n < 21, Kendall's tau-b correlations.

STUDY 3: CALIBRATION STUDY OF MEASURING NON-COGNITIVE FACETS OF COMPETENCE VIA EES

Finally, the computer-based assessment of domain-specific problem-solving competence was implemented in a large-scale study with almost 800 participants in vocational schools in six federal German states. Parts of this final step of the test development are published in Rausch et al. (2016). Hence, parts of the following description are borrowed from Rausch et al. (2016).

Materials and Methods

Participants

A total of 786 VET students participated in the study, of which six were excluded from the analyses due to missing data (due either to lack of willingness or a technical malfunction of the test software). The participating VET students were in the 2nd or 3rd year of a 3-year commercial apprenticeship program, 50.1% were female and the sample showed a typical right skewed age distribution (M = 21.3 years; SD = 2.69; min = 17; max = 44).

Procedure

All data were collected in computer-equipped classrooms in vocational schools. At the beginning of the data collection sessions the researchers introduced the project and the agenda. All participants provided written informed consent. Before and after the problem scenarios, the participants completed several self-report questionnaires including scales on work-related interest and work-related self-concept. In the following, we focus on the internal consistency and internal validity of the assessment of the non-cognitive facets of domain-specific problem-solving competence.

Measures

Embedded experience sampling (EES)

For the main study, four EES events were defined. The first three EES events were the same that were used in the previous pilot study (see **Table 5**: short email response after the reception of the task, phone call from the sender of the task, short visit by a colleague). **Table 8** only lists the additional fourth EES events, the related competence facets, and the EES items that were used.

Generalized self-reports of work-related self-efficacy and work-related interest

We administered a questionnaire on work-related self-efficacy (Abele et al., 2000) which consisted of six statements that had to be rated on a five-point Likert scale ranging from 1 = *disagree* to 5 = *agree* (e.g., "I do not worry about work-related challenges because I can always trust my abilities."). Cronbach's alpha was 0.69. An adapted version of a scale to measure dispositional interests in students (Schiefele et al., 1993) was administered to measure dispositional work-related interest. Six statements had to be rated on a four-point Likert scale ranging from 1 = *disagree* to 4 = *agree* (e.g., "I am sure that I have chosen an apprenticeship program which reflects my personal interests"). Cronbach's alpha was 0.76.

Data Analysis

To assess the cognitive facets of competence (see competence model in **Table 1**), a complex three-step method (similar to Bennett et al., 2003) was applied: (1) Fine-grained results from a highly structured content analysis were condensed into (2) partial credit items on the basis of consensual expert judgments. (3) Finally, these partial credits were subject to psychometric scaling using a multidimensional Rasch model. For further details see Rausch et al. (2016) and Seifried et al. (unpublished).

Results

Requirements of IRT (RQ3)

The variables of non-cognitive facets were calibrated in a six-dimensional partial credit model (Masters, 1982). However, facet D3 ("interest in the progress of/in learning from the problem"), showed insufficient reliability (EAP/PV reliability = 0.30) and therefore was excluded. Thus, the final estimation only included five dimensions and was estimated including background information such as gender, age, vocation, intelligence, competence scores for the cognitive facets, and other relevant variables. All calculations were conducted using the R package TAM (Kiefer et al., 2015). **Table 9** shows the EAP/PV reliabilities (on the diagonal) and the latent correlations between the five remaining non-cognitive competence facets (Rausch et al., 2016).

Correlations With Generalized Retrospective Measures (RQ4)

Furthermore, **Table 9** shows correlations between non-cognitive facets as measured by EES and the corresponding generalized self-report measures of work-related self-efficacy and work-related interest.

Table 9 shows that the EES data meet the requirements of IRT with the exception of D3 (see above). This supports H3b. There are only small correlations between EES-based scores and scores that are based on generalized self-reports, supporting H4c.

TABLE 8 | Additional fourth EES events, competence facets, and EES items in Study 3.

EES event (point of time)	Competence facet (see Table 1)	EES items (translated from German and condensed)
EES event 4: short request from the sender of the task after the reception of the solution (after submission or after 30 min)	Situational confidence in one's solution (C3)	C3: Sender of the task asks how confident the apprentice is about her/his solution and whether the solution has to be checked before its implementation. Answer from 1 = '*Unfortunately*, *I did not arrive at a solution at all*' over 2 = '*I am afraid you should check everything in detail because I assume I made some mistakes*' to 5 = '*I think I found a proper solution that you do not have to check in detail again*.'
	Interest in the progress of/in learning from the problem (D3)	Participants are to check two of the following statements for his email answer. *'Working on tasks like this, . . .* D3_1: . . . *I am always a bit anxious that I might not solve it.'* (*distractor*) D3_2: . . . *I feel as if I am accepted as a full team member.'* (*distractor*) D3_3: . . . *I can always learn something interesting.'* D3_4: . . . *I have the opportunity to demonstrate my skills.'* (*distractor*) D3_5: . . . *I am afraid to make a fool of myself if I fail.'* (*distractor*) D3_6: . . . *I wish that afterwards someone would explain to me how I could have done better.'*

DISCUSSION

Summary of Results

Non-cognitive facets of competence are often neglected in competence assessments. In this paper we introduced Embedded Experience Sampling (EES) as an approach to measuring non-cognitive facets of domain-specific problem-solving competence within a computer-based office simulation. The feasibility and validity of EES were investigated throughout three studies by using different measures and analysis approaches. Most of the results support the validity of EES. The results are discussed with regard to the research questions and hypotheses that were outlined previously (see **Table 2**).

Research question 1 aimed at the test-takers' perception of the EES events in terms of ecological validity. It was hypothesized that participants in low-stake tests do not experience EES events as an additional and unrealistic burden, a finding supported by group discussions and individual interviews in study 1 and by survey data in study 2. Despite experiencing the scenario as quite difficult, they considered it to be authentic and, on average, did not evaluate EES as an additional burden.

TABLE 9 | EAP/PV reliabilities (diagonal) and latent correlations of the non-cognitive facets and generalized self-reports in Study 3.

	(C1)	(C2)	(C3)	(D1)	(D2)
(C1) Situational confidence in one's competence	0.85				
(C2) Ambiguity/uncertainty tolerance	0.57	0.77			
(C3) Situational confidence in one's solution	0.72	0.46	0.84		
(D1) Interest in the problem context/content	0.57	0.62	0.38	0.80	
(D2) Maintaining positive and active emotional states	0.51	0.39	0.45	0.45	0.78
Generalized self-report of work-related self-efficacy	0.29	0.25	0.27	0.16	0.21
Generalized self-report of work-related interest	0.24	0.27	0.15	0.27	0.10

Parts of these results were published in Rausch et al. (2016).

Research question 2 aimed at social desirability as a potential bias in terms of construct validity. In study 1, the participants' responses in group discussions and individual interviews gave no reasons to assume biases from social desirability. In study 2, the dispositional tendency for impression management was uncorrelated with the EES responses and situational impression management (i.e., having thought about which response to the EES items would make someone look good) showed very small correlations with the EES responses. Altogether, social desirability does not appear as a source of bias in EES responses.

Research question 3 aimed at assessing the consistency of the EES data with assumptions of the Multitrait-Multimethod approach (MTMM) and the requirements of a multidimensional model based on Item Response Theory (IRT) in terms of internal validity. In study 1, low correlations of heterotrait-monomethod combinations and higher correlations of monotrait-heteromethod combinations support the assumption of internal validity, however, the differences are only small. In study 3, the EES data was calibrated in a multidimensional IRT model and showed satisfactory EAP/PV reliabilities for five of the six facets while one facet had to be excluded due to low reliability. Altogether, our analysis supports the assumption of internal validity.

Research question 4 aimed at the correlation of EES data with CSSM data (Continuous State Sampling Method) and test motivation in terms of convergent validity and the correlation between EES data and generalized retrospective self-reports in terms of divergent validity. Substantial correlations between EES data and test motivation support the assumption of convergent validity, while the correlations between EES data and CSSM data were more heterogeneous. Low (almost zero) correlations between EES data and generalized retrospective self-reports in study 1 and study 3 emphasize the significance of the measurement approach.

Altogether, we collected data on the feasibility and validity of EES throughout three field studies on problem-solving competence in the business domain and found very promising results. Embedding self-reports on situational experience into the "storyline" of authentic problem scenarios produces reliable and valid data on non-cognitive facets of problem-solving competence.

Limitations and Further Research

Both, the methodological approach and the empirical studies have their limitations. First and foremost, we have not tested for external validity, namely by measuring whether emotional states in the test situations are good proxies for emotional states in respective work situations, which constitutes a strong assumption; not only for the non-cognitive facets but also for the cognitive facets of competence. However, it is very difficult to put together an appropriate research design and collect the respective data to investigate these assumptions. Furthermore, data collection in EES is still based on self-reporting. In our studies, we did not find indications of social desirability or of an additional burden due to EES. However, these studies comprised low-stakes testing. In high-stakes testing, responses to EES items are prone to manipulation and EES events might be experienced as more disruptive. The operationalization of the non-cognitive facets of problem-solving competence is arguable. In study 1, the internal consistencies of EES items measuring the same facet were not satisfactory. Many alternative items would have been just as appropriate or maybe more appropriate as indicators of the respective facet. We have not experimented widely with the operationalization of the facets. One significant alteration concerned the facet D3 "interest in the progress of/in learning from the problem." However, this alteration worsened the model fit and resulted in the exclusion of facet D3 from the IRT model in Study 3, while the correlations within the MTMM analysis in Study 1 had been quite promising. We will vary the item content and the item format in future studies and we encourage other research teams to apply similar approaches in their studies, too.

Limitations of Study 1 and Study 2 were the smaller sample sizes that did not allow for more sophisticated analyses. In Study 2, the CSSM items could have been more similar to the other EES items. Only the items regarding situational interest were very similar. In future studies, more appropriate CSSM items should be applied. Moreover, physiological measures such as heart rate (HR), heart rate variability (HRV), skin conductance or cortisol may be used to further validate the EES data. One such study was conducted by Kärner et al. (2018) who also used the above office simulation and found that CSSM data and physiological data (HR, HRV, and cortisol) showed very similar trends in the course of problem solving. A further data source for validation is the log files from the office simulation. Novak and Johnson (2012) discuss how this non-intrusive data source can be used to measure emotion. Finally, an experimental study in which the participants' emotional experience is manipulated would allow the sensitivity of EES to be tested.

CONCLUSION

Twenty years ago, Weinert (1999) stated that "when assessing competencies, current motivational influences on performance cannot be measured. [...] It is feasible only to measure competence-specific motivational attitudes, for example, with reliable and valid questionnaires" (Weinert, 1999, p. 20). In this paper, we introduced Embedded Experience Sampling (EES) as an alternative method to measure non-cognitive facets of competence within the performance assessment instead of relying on decontextualized general self-reports. The idea behind EES is that the repeated measurement of emotional or motivational states during domain-specific tasks allows for an inference to be made regarding non-cognitive traits; similar to Chomsky (1965) distinction between manifest performance and latent competence. This helps to overcome the asymmetry in the content and breadth in the measurement of the cognitive and non-cognitive constructs (Brunswik, 1956).

Drawing on our experience, EES is a feasible and informative approach to measuring non-cognitive facets of competence under the following conditions: (1) The computer-based performance assessment is embedded in an immersive and authentic simulation of a real-life domain. (2) The participants are confronted with comprehensive scenarios that require a sustained performance. (3) The participants are introduced to EES within a tutorial prior to the performance assessment. Drawing on our empirical studies, we found indications of the validity of EES. We would like to encourage other researchers to implement EES or similar approaches into their studies of competence assessment because further research is needed for the subsequent development and validation of the method.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the Ethical Board of the University of Bamberg, Germany and the Board of Data Protection of the Federal State Authority of Bavaria (Germany) with written informed consent from all subjects.

AUTHOR CONTRIBUTIONS

AR, KK, and JS contributed to method, data collection, and preparation of the manuscript. AR led the project.

FUNDING

This research was funded by the German Federal Ministry of Education and Research und Grant No. 01DB081119–01DB1123. The publication of this article was funded by the Ministry of Science, Research and the Arts Baden-Württemberg and the University of Mannheim.

ACKNOWLEDGMENTS

The authors would like to thank the DomPL-IK research group: Christin Siegfried, Detlef Sembill, Eveline Wuttke, Jan Küster, Karsten D. Wolf, Marc Egloffstein, Rebecca Eigenmann, Steffen Brandt, and Thomas Schley.

REFERENCES

Abele, A. E., Stief, M., and Andrä, M. (2000). Zur ökonomischen Erfassung beruflicher Selbstwirksamkeitserwartungen – Neukonstruktion einer BSW-Skala. *Zeitschrift für Arbeits- und Organisationspsychologie* 48, 4–16.

Bennett, R. E., Jenkins, F., Persky, H., and Weiss, A. (2003). Assessing complex problem solving performances. *Assess. Educ.* 10, 347–359. doi: 10.1016/j.jecp.2017.04.011

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments*. Berkeley, CA: University of California Press.

Brynjolfsson, E., and McAfee, A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York, NY: W. W. Norton & Company.

Butler, J., and Adams, R. J. (2007). The impact of differential investment of student effort on the outcome of international studies. *J. Appl. Meas.* 3, 279–304.

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56, 81–105. doi: 10.1037/h0046016

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edn. Hillsdale, NJ: Lawrence Erlbaum Associates.

Conrad, M., and Schumann, S. (2017). Lust und Frust im Tablet-PC-basierten Wirtschaftsunterricht – Befunde einer Interventionsstudie zur Erfassung des affektiven Unterrichtserlebens mittels Continuous-State-Sampling. *Zeitschrift für Berufs- und Wirtschaftspädagogik* 113, 33–55.

Cortina, J. M., and Luchman, J. N. (2012). "Personnel selection and employee performance," in *Handbook of Psychology*, 2nd Edn, eds I. B. Weiner, N. W. Schmitt, and S. Highhouse (Hoboken, NJ: Wiley), 143–183.

Cronbach, L. J. (1960). *Essentials of Psychological Testing*, 2nd Edn. New York, NY: Harper & Row.

Csikszentmihalyi, H., and Larson, R. (1987). Validity and reliability of the experience sampling method. *J. Nerv. Mental Dis.* 175, 526–536. doi: 10.1097/00005053-198709000-00004

Debus, G. (2000). "Sprachliche Methoden [Verbal methods]," in *Emotionspsychologie – Ein Handbuch [Emotion Psychology – A Handbook]*, eds J. Otto, H. Euler, and H. Mandl (Weinheim: Psychologie Verlags Union), 409–418.

Dermitzaki, I., Leondari, A., and Goudas, M. (2009). Relations between young students' strategic behaviours, domain-specific self-concept, and performance in a problem-solving situation. *Learn. Instr.* 19, 144–157. doi: 10.1016/j.learninstruc.2008.03.002

Diener, E., and Lucas, R. (2000). "Subjective emotional well-being," in *Handbook of Emotions*, eds M. Lewis and J. Haviland- Jones (New York: Guilford Press), 325–337.

Dörner, D., and Funke, J. (2017). Complex problem solving: what it is and what it is not. *Front. Psychol.* 8:1153. doi: 10.3389/fpsyg.2017.01153

Drechsel, B., Carstensen, C., and Prenzel, M. (2011). The role of content and context in PISA interest scales: a study of the embedded interest items in the PISA 2006 science assessment. *Int. J. Sci. Educ.* 33, 73–95. doi: 10.1080/09500693.2010.518646

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., and Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proc. Natl. Acad. Sci. U.S.A.* 108, 7716–7720. doi: 10.1073/pnas.1018601108

Duckworth, A. L., and Yeager, D. S. (2015). Measurement matters: assessing personal qualities other than cognitive ability for educational purposes. *Educ. Res.* 44, 237–251. doi: 10.3102/0013189x15584327

Eklöf, H. (2010). Skill and will: test-taking motivation and assessment quality. *Assess. Educ.* 17, 345–356. doi: 10.1080/0969594x.2010.516569

Frensch, P. A., and Funke, J. (1995). "Definitions, traditions, and a general framework for understanding complex problem solving," in *Complex Problem Solving: The European perspective*, eds P. A. Frensch and J. Funke (Hillsdale: Lawrence Erlbaum Associates), 3–25.

Frey, C. B., and Osborne, M. A. (2017). The future of employment: how susceptible are jobs to computerisation? *Technol. Forecast. Soc. Change* 114, 254–280. doi: 10.1016/j.techfore.2016.08.019

Hannula, M. S. (2015). "Emotions in problem solving," in *Selected Regular Lectures from the 12th International Congress on Mathematical Education*, ed. S. J. Cho (New York, NY: Springer), 269–288. doi: 10.1007/978-3-319-17187-6_16

Herl, H. E., O'Neil, H. F., Chung, G. K., Bianchi, C., Wang, S., Mayer, R., et al. (1999). *Final Report For Validation Of Problem-Solving Measures*. Los Angeles, CA: University of California. .

Isen, A. M. (2008). "Some ways in which positive affect influences decision making and problem solving," in *Handbook of Emotions*, eds M. Lewis, J. M. Haviland-Jones, and L. Feldman Barrett (New York: Guilford Press), 548–573.

Kanfer, R., and Ackerman, P. L. (2005). "Work competence. a person-oriented perspective," in *Handbook of Competence and Motivation*, eds A. J. Elliot and C. S. Dweck (New York: Guilford Press), 336–353.

Kärner, T., Minkley, N., Rausch, A., Schley, T., and Sembill, D. (2018). Stress and resources in vocational problem solving. *Vocations and Learning* 11, 365–398. doi: 10.1007/s12186-017-9193-8

Kärner, T., Sembill, D., Aßmann, C., Friederichs, E., and Carstensen, C. H. (2017). Analysis of person-situation interactions in educational settings via cross-classified multilevel longitudinal modelling: illustrated with the example of students' stress experience. *Front. Learn. Res.* 5, 16–42. doi: 10.14786/flr.v5i1.137

Kiefer, T., Robitzsch, A., and Wu, M. (2015). *TAM: Test Analysis Modules (version 1.3)*. Available at: https://cran.r-project.org/web/packages/TAM/index.html (accessed March 18, 2019).

Kögler, K., and Göllner, R. (2018). Control-value appraisals predicting students' boredom in accounting classes: a continuous-state-sampling approach. *Empir. Res. Vocat. Educ. Train.* 10:4.

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., et al. (2002). *German Scale Handbook for PISA 2000*. Berlin: Max-Planck-Institut für Bildungsforschung.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika* 47, 149–174. doi: 10.1007/bf02296272

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educ. Res.* 23, 13–23. doi: 10.3102/0013189x023002013

Musch, J., Brockhaus, R., and Bröder, A. (2002). Ein Inventar zur Erfassung von zwei Faktoren sozialer Erwünschtheit. *Diagnostica* 48, 121–129. doi: 10.1026//0012-1924.48.3.121

Neisser, U. (1967). *Cognitive Psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Adv. Health Sci. Educ.* 15, 625–632. doi: 10.1007/s10459-010-9222-y

Novak, E., and Johnson, T. E. (2012). "Assessment of student's emotions game-based learning in," in *Assessment in Game-Based Learning: Foundations, Innovations and Perspectives*, eds D. Ifenthaler, D. Eseryel, and X. Ge (New York, NY: Springer), 379–399. doi: 10.1007/978-1-4614-3546-4_19

Paulhus, D. L. (1994). *Balanced Inventory of Desirable Responding: Reference Manual for BIDR Version 6*. Vancouver: University of British Columbia.

Pekrun, R. (2006). The Control-Value theory of achievement emotions: assumptions, corollaries, and implications for educational research and practice. *Educ. Psychol. Rev.* 18, 315–341. doi: 10.1007/s10648-006-9029-9

Pellegrino, J. W., DiBello, L. V., and Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educ. Psychol.* 51, 59–81. doi: 10.1080/00461520.2016.1145550

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., and Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *J. Appl. Psychol.* 88, 879–903. doi: 10.1037/0021-9010.88.5.879

Rausch, A. (2017). Dispositional predictors of problem solving in the field of office work. *Vocat. Learn.* 10, 177–199. doi: 10.1007/s12186-016-9165-4

Rausch, A., Seifried, J., Wuttke, E., Kögler, K., and Brandt, S. (2016). Reliability and validity of computer-based assessment of cognitive and non-cognitive facets of problem-solving competence in the business domain. *Empir. Res. Vocat. Educ. Train.* 8, 1–23

Rausch, A., and Wuttke, E. (2016). Development of a multi-faceted model of domain-specific problem-solving competence and its acceptance by different stakeholders in the business domain. *Unterrichtswissenschaft* 44, 164–189.

Reis, H. T. (2012). "Why researchers should think "real-world": a conceptual rationale," in *Handbook of Research Methods for Studying Daily Life*, eds M. R. Mehl and T. S. Conner (New York, NY: Guilford Press), 3–21.

Robinson, M. D., and Clore, G. L. (2002). Episodic and semantic knowledge in emotional self-report: evidence for two judgment processes. *J. Pers. Soc. Psychol.* 83, 198–215. doi: 10.1037//0022-3514.83.1.198

Rychen, D. S., and Salganik, L. H. (2003). "A holistic model of competence," in *Key Competencies for a Successful Life and Well-Functioning Society*, eds D. S. Rychen and L. H. Salganik (Göttingen: Hogrefe & Huber), 41–62.

Sabourin, J. L., and Lester, J. C. (2014). Affect and engagement in game-based learning environments. *IEEE Trans. Affect. Comput.* 5, 45–56. doi: 10.1109/T-AFFC.2013.27

Schiefele, U., Krapp, A., Wild, K.-P., and Winteler, A. (1993). Der "Fragebogen zum Studieninteresse" (FSI) [The 'questionnaire on students' interest' (FSI)]. *Diagnostica* 39, 335–351.

Schoenfeld, A. H. (2013). Reflections on problem solving theory and practice. *Math. Enthusiast* 10, 9–34.

Schoppek, W., and Fischer, A. (2015). Complex problem solving— single ability or complex phenomenon? *Front. Psychol.* 6:1669. doi: 10.3389/fpsyg.2015.01669

Schwarz, N. (2012). "Why researchers should think "real-time": a cognitive rationale," in *Handbook of Research Methods for Studying Daily Life*, eds M. R. Mehl and T. S. Conner (New York, NY: Guilford Press), 22–42.

Scollon, C. N., Kim-Prieto, C., and Diener, E. (2003). Experience-sampling: promises and pitfalls, strengths and weaknesses. *J. Happiness Stud.* 4, 5–34. doi: 10.1023/A:1023605205115

Sembill, D. (1992). *Problemlösefähigkeit, Handlungskompetenz und Emotionale Befindlichkeit – Zielgrößen forschenden Lernens [Problem-Solving Ability, Action Competence, and Emotional State]*. Göttingen: Hogrefe.

Sembill, D., Rausch, A., and Kögler, K. (2013). "Non-cognitive facets of competence. Theoretical foundations and implications of measurement," in *From Diagnostics to Learning Success. Proceedings in Vocational Education and Training*, eds K. Beck and O. Zlatkin-Troitschanskaia (Rotterdam: Sense), 199–212.

Sembill, D., Seifried, J., and Dreyer, K. (2008). PDAs als Erhebungsinstrument in der beruflichen Lernforschung – Ein neues Wundermittel oder bewährter Standard? *Empirische Pädagogik* 22, 64–77.

Stodel, M. (2015). But what will people think?: getting beyond social desirability bias by increasing cognitive load. *Int. J. Market Res.* 57, 313–321.

Sugrue, B. (1995). A theory-based framework for assessing domain-specific problem-solving ability. *Educ. Meas.* 14, 29–36.

van Reekum, C. M., and Scherer, K. R. (1997). "Levels of processing in emotion antecedent appraisal," in *Cognitive Science Perspectives on Personality and Emotion*, ed. G. Matthews (New York, NY: Elsevier), 259–300. doi: 10.1016/s0166-4115(97)80123-9

Verschaffel, L., Dooren, W. V., and De Smedt, B. (2012). "Mathematical learning," in *Encyclopedia of the Sciences of Learning*, ed. N. M. Seel (New York, NY: Springer), 2107–2110.

Weinert, F. E. (1999). *Concepts of Competence. Expert Report for the OECD Project Definition and Selection of Competencies (DeSeCo)*. Munich: Max Planck Institute for Psychological Research.

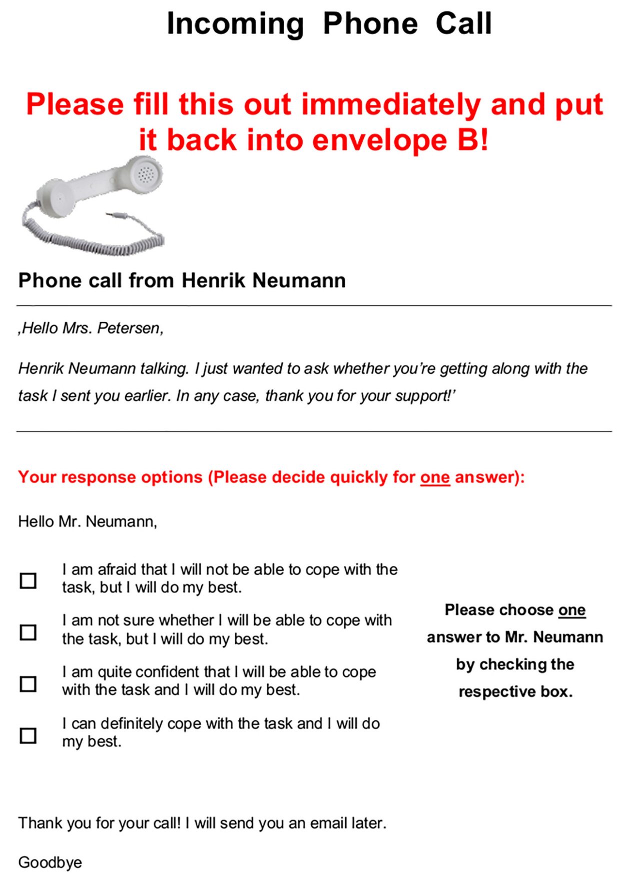
Weinert, F. E. (2001). "Concept of competence: a conceptual clarification," in *Defining and Selecting Key Competencies*, eds D. S. Rychen and L. H. Salganik (Seattle: Hogrefe and Huber), 45–65.

Wigfield, A. and Eccles, J. S. (2000). Expectancy-Value theory of achievement motivation. *Contemp. Educ. Psychol.* 25, 68–81. doi: 10.1006/ceps.1999.1015

Wittmann, W. W., and Süß, H.-M. (1999). "Investigating the paths between working memory, intelligence, knowledge, and complex problem-solving performances via brunswik symmetry," in *Learning and Individual Differences: Process, Trait, and Content Determinants*, eds P. L. Ackerman, P. C. Kyllonen, and R. D. Roberts (Washington, DC: American Psychological Association), 77–108.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX



Incoming Phone Call

Please fill this out immediately and put it back into envelope B!

Phone call from Henrik Neumann

‚Hello Mrs. Petersen,

Henrik Neumann talking. I just wanted to ask whether you're getting along with the task I sent you earlier. In any case, thank you for your support!'

Your response options (Please decide quickly for one answer):

Hello Mr. Neumann,

□ I am afraid that I will not be able to cope with the task, but I will do my best.

□ I am not sure whether I will be able to cope with the task, but I will do my best.

□ I am quite confident that I will be able to cope with the task and I will do my best.

□ I can definitely cope with the task and I will do my best.

Please choose one answer to Mr. Neumann by checking the respective box.

Thank you for your call! I will send you an email later.

Goodbye

FIGURE A1 | Paper-based EES event "Phone call" with one EES item (translated from German).

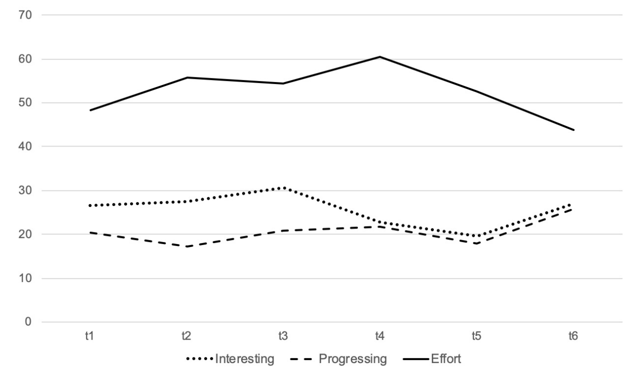


FIGURE A2 | Mean scores of the CSSM items for each measurement point.

TABLE A1 | Descriptive statistics of EES items in Study 1.

	Scenario 1		Scenario 2		Scenario 3	
EES variable	***M***	***SD***	***M***	***SD***	***M***	***SD***
C1 (Situational confidence in one's competence)	2.16	0.57	2.72	0.56	1.77	0.65
C3 (Situational confidence in one's solution)	2.22	0.88	2.97	0.65	1.71	0.96
D2 (Maintaining positive and active emotional states)	2.48	0.61	3.19	0.46	2.32	0.65
D3 (Interest in the progress of/in learning from the problem)	1.87	1.06	2.10	1.20	1.83	1.06

*See **Table 3** for corresponding EES items.*

TABLE A2 | Descriptive statistics of EES items in Study 2.

	Range	Minimum	Maximum	*M*	*SD*
EES variable C1	1–4	1.00	4.00	1.58	0.71
EES variable C2	1–4	1.00	4.00	2.15	0.81
EES variable D1	1–4	1.00	2.00	1.57	0.51
EES variable D2	1–4	1.00	3.25	2.19	0.64
CSSM scale Interesting	1–100	3.83	49.33	24.34	13.62
CSSM scale Effort	1–100	13.17	89.17	52.06	22.34
CSSM scale Progress	1–100	1.17	41.17	19.80	13.61
Impression management (BIDR scale)	1–4	1.14	3.43	2.42	0.56
Impression management (single item)	1–5	1.00	4.00	1.80	0.89
Authenticity of EES events	1–5	1.00	5.00	3.65	1.27
Additional burden of EES events	1–5	1.00	5.00	2.45	1.23
Test motivation (Effort Thermometer)	1–10	3.00	10.00	6.56	2.55

*See **Table 5** for corresponding EES items.*

ORIGINAL RESEARCH
published: 06 June 2019
doi: 10.3389/fpsyg.2019.01277

Evaluating Different Equating Setups in the Continuous Item Pool Calibration for Computerized Adaptive Testing

Sebastian Born[1]*, Aron Fink[2], Christian Spoden[3] and Andreas Frey[2,4]

[1] Department of Research Methods in Education, Institute of Educational Science, Friedrich Schiller University Jena, Jena, Germany, [2] Educational Psychology: Measurement, Evaluation and Counseling, Institute of Psychology, Goethe University Frankfurt, Frankfurt, Germany, [3] German Institute for Adult Education, Leibniz Centre for Lifelong Learning, Bonn, Germany, [4] Faculty of Educational Sciences, Centre for Educational Measurement, University of Oslo, Oslo, Norway

OPEN ACCESS

Edited by:
Ronny Scherer,
University of Oslo, Norway

Reviewed by:
Alvaro J. Arce-Ferrer,
Pearson, United States
Alexander Robitzsch,
Christian-Albrechts-Universität zu Kiel,
Germany

***Correspondence:**
Sebastian Born
sebastian.born@uni-jena.de

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: 23 November 2018
Accepted: 15 May 2019
Published: 06 June 2019

Citation:
Born S, Fink A, Spoden C and
Frey A (2019) Evaluating Different
Equating Setups in the Continuous
Item Pool Calibration
for Computerized Adaptive Testing.
Front. Psychol. 10:1277.
doi: 10.3389/fpsyg.2019.01277

The increasing digitalization in the field of psychological and educational testing opens up new opportunities to innovate assessments in many respects (e.g., new item formats, flexible test assembly, efficient data handling). In particular, computerized adaptive testing provides the opportunity to make tests more individualized and more efficient. The newly developed continuous calibration strategy (CCS) from Fink et al. (2018) makes it possible to construct computerized adaptive tests in application areas where separate calibration studies are not feasible. Due to the goal of reporting on a common metric across test cycles, the equating is crucial for the CCS. The quality of the equating depends on the common items selected and the scale transformation method applied. Given the novelty of the CCS, the aim of the study was to evaluate different equating setups in the CCS and to derive practical recommendations. The impact of different equating setups on the precision of item parameter estimates and on the quality of the equating was examined in a Monte Carlo simulation, based on a fully crossed design with the factors common item difficulty distribution (bimodal, normal, uniform), scale transformation method (mean/mean, mean/sigma, Haebara, Stocking-Lord), and sample size per test cycle (50, 100, 300). The quality of the equating was operationalized by three criteria (proportion of feasible equatings, proportion of drifted items, and error of transformation constants). The precision of the item parameter estimates increased with increasing sample size per test cycle, but no substantial difference was found with respect to the common item difficulty distribution and the scale transformation method. With regard to the feasibility of the equatings, no differences were found for the different scale transformation methods. However, when using the moment methods (mean/mean, mean/sigma), quite extreme levels of error for the transformation constants *A* and *B* occurred. Among the characteristic curve method the performance of the Stocking-Lord method was slightly better than for the Haebara method. Thus, while no clear recommendation can be made with regard to the common item difficulty distribution, the characteristic curve methods turned out to be the most favorable scale transformation methods within the CCS.

Keywords: computerized adaptive test, item response theory, equating, continuous calibration, simulation

INTRODUCTION

The shift to using digital technology (e.g., laptops, tablets, and smartphones) for psychological and educational assessments provides the opportunity to implement computer-based state-of-the-art methods from psychometrics and educational measurement in day-to-day testing practice. In particular, computerized adaptive testing (CAT) has the potential to make tests more individualized and to enhance efficiency (e.g., Segall, 2005). CAT is a method of test assembly that uses the responses given to previously presented items for the selection of the next item (e.g., van der Linden, 2016), whereby the item that satisfies a statistical optimality criterion best is selected from a precalibrated item pool. Therefore, the calibrated item pool is an essential and important building block in CAT (e.g., Thompson and Weiss, 2011; He and Reckase, 2014). A set of items is called a calibrated item pool if the item characteristics, such as item difficulty and item discrimination, were estimated on the basis of an item response theory (IRT; e.g., van der Linden, 2016) model beforehand. However, in some contexts, such as higher education, clinical diagnosis, or personnel selection, the item pool calibration for CAT often poses a critical challenge because separate calibration studies are not feasible, and sample sizes are too low to allow for stable item parameter estimation.

To overcome this problem, Fink et al. (2018) proposed a continuous calibration strategy (CCS), which enables a step-by-step build-up of the item pool across several test cycles during the operational CAT phase. In the context of the CCS a test cycle is understood as the whole test procedure including steps like test assembly, test administration and analysis of test results. As the item parameter estimates of existing and new items are continuously updated within the CCS, equating is a critical factor to enable interchangeable score interpretation across test cycles. The equating procedure implemented in the CCS is based on a common-item non-equivalent group design (Kolen and Brennan, 2014) and is carried out in four steps: (1) common item selection, (2) scale transformation, (3) item parameter drift (IPD; e.g., Goldstein, 1983) detection, and (4) fixed common item parameter (FCIP; e.g., Hanson and Béguin, 2002) calibration.

In their study, Fink et al. (2018) evaluated the performance of the CCS for different factors (sample size per test cycle, calibration speed, and IRT model) with respect to the quality of the person parameter estimates. Although the results were promising, two issues remained open. First, the study of Fink et al. (2018) was conducted under ideal conditions (i.e., constant ability distribution of the examinees across test cycles). Second, despite the importance of the equating procedure in the CCS, its performance with respect to different setups of the procedure (i.e., selection of common items, scale transformation method, item drift detection) was not investigated in detail. For example, it became apparent that the CCS did not work as intended for very easy or very difficult items when using small sample sizes (i.e., 50 or 100 examinees) per test cycle. In these cases, item parameter estimates were biased due to a few inconsistent responses, with the consequence that these items were no longer selected by the adaptive algorithm in the following test cycles. Therefore, it was not possible to continuously update the item parameter estimates for these items.

Against this background, the aim of the present study was to investigate the performance of the equating procedure for different setups conducted under more realistic conditions (i.e., examinees' average abilities and variance differ between test cycles). The remainder of the article is organized as follows: First, we provide the theoretical background for the present study by introducing the underlying IRT model and by describing the CCS. Next, we discuss both the previously implemented equating procedure and alternative specifications. Then, we examine the performance of different setups of the different equating procedures in a simulation. Finally, we discuss the results and make recommendations for the implementation of the CCS.

THEORETICAL BACKGROUND

IRT Model

The IRT model used in this study was the two-parameter logistic (2PL) model (Birnbaum, 1968) for dichotomous items. The 2PL model defines the probability of a correct response $u_{\mathrm{ij}} = 1$ of examinee $j = 1 \ldots N$ with a latent ability level θ_{j} to an item i by the following model, whereby a_{i} is the discrimination parameter and d_{i} is the easiness parameter of item i:

$$P\left(u_{\mathrm{ij}} = 1|\theta_{\mathrm{j}}, a_{\mathrm{i}},\ d_{\mathrm{i}}\right) = \frac{\exp\left(a_{\mathrm{i}}\theta_{\mathrm{j}} + d_{\mathrm{i}}\right)}{1 + \exp\left(\ a_{\mathrm{i}}\theta_{\mathrm{j}} + d_{\mathrm{i}}\right)}, \tag{1}$$

In the traditional IRT metric where $a_{\mathrm{i}}\theta_{\mathrm{j}} + d_{\mathrm{i}} = \ a_{\mathrm{i}}\left(\theta_{\mathrm{j}} - b_{\mathrm{i}}\right)$, the a_{i} parameters will be the identical for these parametrizations, while the item difficulty parameter b_{i} is calculated as $b_{\mathrm{i}} = -d_{\mathrm{i}}\ /a_{\mathrm{i}}$.

Continuous Calibration Strategy

In the following paragraphs, we briefly outline the CCS as introduced by Fink et al. (2018) and detail the equating procedure implemented. The CCS consists of two phases, a non-adaptive *initial phase* and a partly adaptive *continuous phase*. In the initial phase, which is the first test cycle of the CCS, the same items are presented to all examinees and only the item order can vary between examinees. In the continuous phase, the tests assembled consist of three types of item clusters (calibration cluster, linking cluster, adaptive cluster), whereby a cluster is comprised of several items. Each type of cluster has a specific goal. The calibration cluster offers the opportunity to include new items in the existing item pool, the linking cluster utilizes common items to allow a scale to be established across test cycles, and the adaptive cluster aims at the enhancement of measurement precision. The items in the calibration and the linking clusters are the same for all examinees and are administered sequentially, whereas the items in the adaptive cluster can differ between examinees due to the adaptive selection algorithm. Each test cycle in the continuous phase can be broken down into seven steps: (1) common item selection for the linking cluster, (2) test assembly and test administration, (3) temporary item parameter estimation, (4) scale transformation of the common items, (5) IPD detection for the common items, (6) FCIP calibration, and (7) person parameter estimation. The equating procedure consists of four of these steps, which will be detailed in the following four paragraphs. The first three steps of the equating procedure serve as quality assurance of the common items to ensure feasible equating in the fourth step.

In the *common item selection*, items that have already been calibrated in the previous test cycles are selected as common items for the linking cluster. To ensure that the common items represent the statistical characteristics of the item pool (Kolen and Brennan, 2014), such as the range of the item difficulty, the items are assigned to five categories (very low, low, medium, high, and very high) based on their easiness parameters d_i. Fink et al. (2018) selected the items from the categories in such a way that the difficulty distribution of the common items corresponded approximately to a normal distribution. Beside the representation of the statistical item pool characteristics it is important that the common items adequately reflect the content of the item pool. This can be done by using content balancing approaches (e.g., van der Linden and Reese, 1998; Cheng and Chang, 2009; Born and Frey, 2017) within the common item selection and within the adaptive cluster.

After test assembly and test administration, the parameters for the common items are estimated based on the responses of the current test cycle. In the second step of the equating procedure, a *scale transformation* of the common items has to be conducted, because the ability distribution of the examinees usually differs between test cycles and, therefore, the item parameter estimates obtained are not directly comparable across cycles. The comparability of the parameter estimates is a necessary condition to check whether the common items are affected by IPD. For this reason, scale transformation methods (e.g., Marco, 1977; Haebara, 1980; Loyd and Hoover, 1980; Stocking and Lord, 1983) are important for the equating procedure. Fink et al. (2018) used the mean/mean method (Loyd and Hoover, 1980) for the scale transformation.

As IPD of item parameters may have a serious impact on equating results such as scaled scores and passing rates (Hu et al., 2008; Miller and Fitzpatrick, 2009), the *IPD detection* as the third step of the equating procedure is important if the method is to operate optimally. A number of tests for IPD can be used in IRT-based equating procedures, such as the Lord's χ^2-test (Lord, 1980) and the likelihood-ratio test (Thissen et al., 1988). In an iterative process of scale transformation and testing for IPD, common items that show significant IPD are excluded from the final set of common items. The iterative purification continues as long as at least one of the remaining common items shows significant IPD or less than two common items are left. The rationale behind the latter stopping rule is that at least two link items are necessary to keep the scale comparable across test cycles. Nevertheless, it should be mentioned that with a smaller number of link items, the equating procedure is more prone to sampling errors (Wingersky and Lord, 1984). Fink et al. (2018) used a one-sided t-test to examine whether the parameter estimates of a common item from the current test cycle differed significantly from the parameter estimates of the same item from the preceding test cycle.

The last step of the equating procedure, the *FCIP calibration*, involves the parameter estimation of all items using marginal maximum likelihood (MML; Bock and Aitkin, 1981) based on the responses from all test cycles. Because one aim of the CCS is to maintain the original scale from the initial calibration (first test cycle), the use of one step procedures (e.g., concurrent calibration; Wingersky and Lord, 1984) for estimating all item parameters of the different test cycles in one run is not suitable. If maintaining the scale from the initial calibration over the following test cycles has no priority, promising methods exist for equating multiple test forms simultaneously (Battauz, 2018). In the FCIP calibration, the parameters of the final common items are fixed at the item parameters estimated from the previous test cycle, whereas all the other items are estimated freely. If a "breakdown" occurs, which means that less than two common items remain after the IPD detection, a concurrent calibration (Wingersky and Lord, 1984) is used to establish a new scale.

Specifications of the Common Item Selection

The common item selection and the scale transformation of the common items are crucial parts of the CCS because they ensure that the procedure functions well. In terms of the common item selection, different distributional assumptions such as an approximated normal distribution, as used in Fink et al. (2018), or a uniform distribution may underlie the item selection. Up to now, only Vale et al. (1981) examined the impact of different common item distributions on the accuracy of the item parameter estimates using the mean/sigma method (Marco, 1977). The authors selected the common items in such a way that the test information curves of the common items were peaked (with the most information at theta equals zero) or had an approximately normal or uniform shape. In terms of the bias of the item parameter estimates, the peaked test information curve performed worst. There were only slight differences in the performance, depending on whether normally or uniformly shaped test information curves were used for the common items. As an alternative, items with extreme difficulties (bimodal distribution) might be selected as common items for the linking cluster and, therefore, might be administered to all examinees. As a consequence, the number of responses for these items increases and the impact of the few inconsistent responses that might cause bias in the estimates and prevent later administration and parameter updating in the following test cycles would be reduced. Because the quality of the equating highly depends on the common items selected, it may be argued that especially a bimodal distribution of the common items threatens the goal of maintaining the scale across test cycles. However, the item drift test implemented in the CCS ensures that significant changes in the parameter estimates of the common items between test cycles do not affect the later FCIP calibration that is used to maintain the scale.

Scale Transformation

When item parameters are estimated using different groups of examinees, the obtained parameters are often not comparable

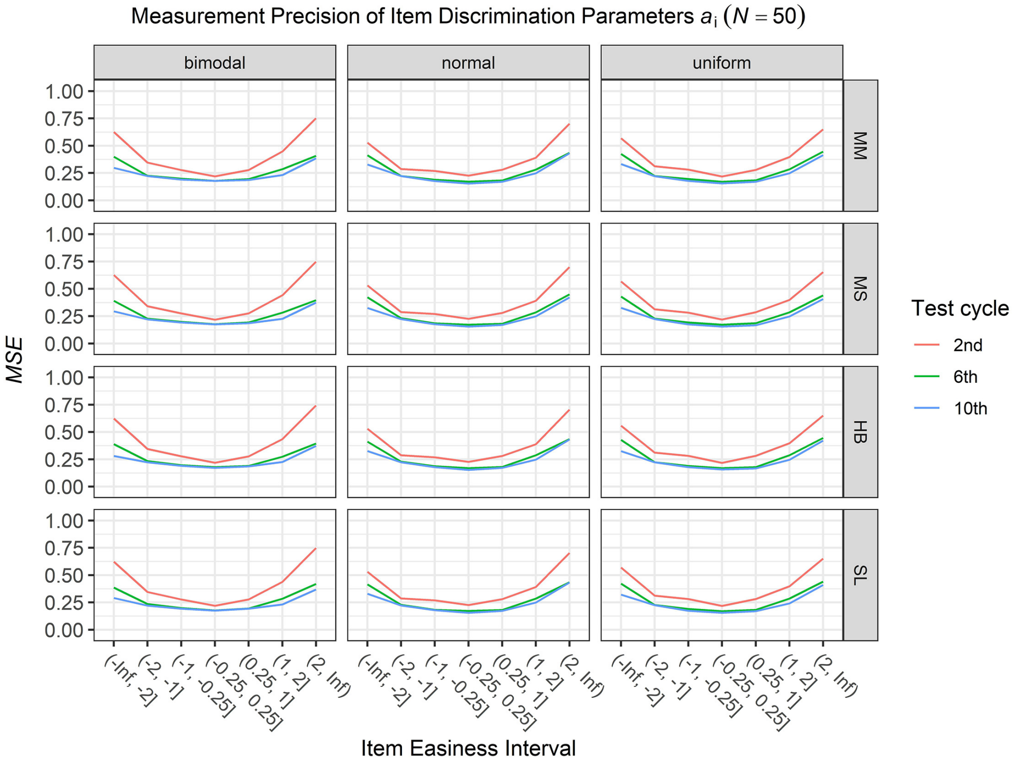


FIGURE 1 | Conditional mean squared error (*MSE*) of the item discrimination a_i for specific item easiness intervals after the 2nd, 6th, and 10th test cycles in the continuous calibration strategy with a sample size per test cycle of $N = 50$ for different common item difficulty distributions and different scale transformation methods (MM = Mean/Mean, MS = Mean/Sigma, HB = Haebara, SL = Stocking-Lord).

due to arbitrary decisions that have been made to fix the scale of the item and person parameter space (Yousfi and Böhme, 2012). In that case, the comparability of the item parameters can be attained by an IRT scale transformation. If the underlying IRT model holds for two groups of examinees, *K* and *L*, then the logistic IRT scales differ by a linear transformation for both the item parameters and the person parameters (Kolen and Brennan, 2014). The linear equation for the θ-values can be formulated as follows:

$$\theta_{\mathrm{Lj}} = A\theta_{\mathrm{Kj}} + B, \qquad (2)$$

where *A* and *B* represent the transformation constants (also referred to as slope and shift) and θ_{Kj} and θ_{Lj} the person parameter values for an examinee *j* on scale *K* and scale *L*. The item parameters for the 2PL model on the two scales are defined in Eqs 3 and 4, where a_{Ki}, b_{Ki}, and a_{Li}, b_{Li} represent the item parameters on scale *K* and on scale *L*, respectively.

$$a_{\mathrm{Li}} = \frac{a_{\mathrm{Ki}}}{A} \qquad (3)$$

$$b_{\mathrm{Li}} = Ab_{\mathrm{Ki}} + B \qquad (4)$$

To obtain the transformation constants *A* and *B*, several scale transformation methods can be used. The *moment methods* such as the mean/mean and the mean/sigma express the relationship of scales by using the means and standard deviations of item or person parameters, whereas the *characteristic curve methods* minimize a discrepancy function with respect to the item characteristic curves (Haebara, 1980) or the test characteristic curve (Stocking and Lord, 1983). Research comparing these methods has found that characteristic curve methods produced more stable results compared to the moment methods (e.g., Baker and Al-Karni, 1991; Kim and Cohen, 1992; Hanson and Béguin, 2002). Within the moment methods, the mean/mean method turned out to be more stable (Ogasawara, 2000). Furthermore, Kaskowitz and de Ayala (2001) found that characteristic curve methods were robust against moderate estimation errors and were more accurate with a larger number of common items (15 or 25 compared to only five common items). In sum, the moment methods are easily implementable, but the characteristic curve methods seem to be more robust against estimation errors.

RESEARCH QUESTIONS

As the purpose of equating procedures in the CCS is to enable an interchangeable score interpretation across test cycles, the selection of the common items is a crucial factor for feasible equating. Up to now, only recommendations for the number of common items that should be used when conducting IRT equating have been made (Kolen and Brennan, 2014). Furthermore, it is suggested that the common items should

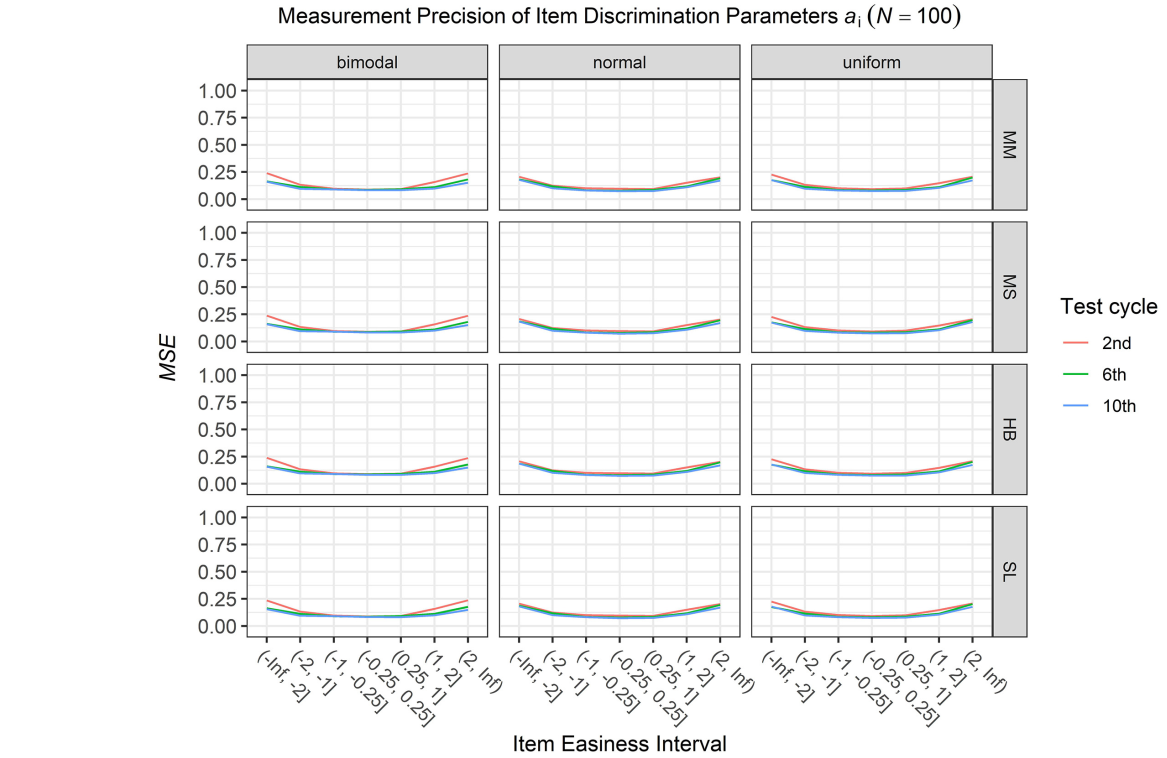


FIGURE 2 | Conditional mean squared error (*MSE*) of the item discrimination a_i for specific item easiness intervals after the 2nd, 6th, and 10th test cycle in the continuous calibration strategy with a sample size per test cycle of N = 100 for different common item difficulty distributions and different scale transformation methods (MM = Mean/Mean, MS = Mean/Sigma, HB = Haebara, SL = Stocking-Lord).

represent the content and statistical characteristics of the test or rather the complete item pool. For example, modifying the common item selection in such a way that more items with extreme item difficulty levels are included may enhance the precision of these items, but it could threaten the quality of the equating. Therefore, our first two research questions can be formulated as follows:

1. What effect does the difficulty distribution of the common items in the CCS have on the precision of the item parameter estimates?
2. What effect does the difficulty distribution of the common items in the CCS have on the quality of the equating?

Fink et al. (2018) used the mean/mean method for scale transformation because of its simple and user-friendly implementation. Given prior research on scale transformation methods, this might not be the best choice when the sample size per test cycle is low. Furthermore, there are several packages for the open-source software R (R Core Team, 2018) available to implement the characteristic curve methods (e.g., Weeks, 2010; Battauz, 2015). As already mentioned above, the scale transformation method used and the IPD detection implemented in the CCS could serve as quality assurance to ensure that significant changes in the parameter estimates of the common items between test cycles do not affect the later FCIP calibration. For this reason, our third research question is:

3. What effect does the scale transformation method used in the CCS have on the quality of the equating?

As the CCS was developed for a context in which separate calibration studies are often not feasible and sample sizes are too low to allow for stable item parameter estimation, it is important to evaluate whether the results for these three research questions were affected by the sample size. Consequently, each of the three research questions was investigated with a special focus on additional variations of the sample size.

MATERIALS AND METHODS

Study Design

Many factors can affect the quality of the equating within the CCS. These include, among others, the number of common items, the test length, the characteristics of the common items, the scale transformation method applied, the number of examinees per test cycle, the presence of IPD and the test applied for IPD. In the present study, some of these factors were kept constant (e.g., number of common items, test length, the presence of IPD, test applied for IPD) to ensure the comprehensibility of the study results.

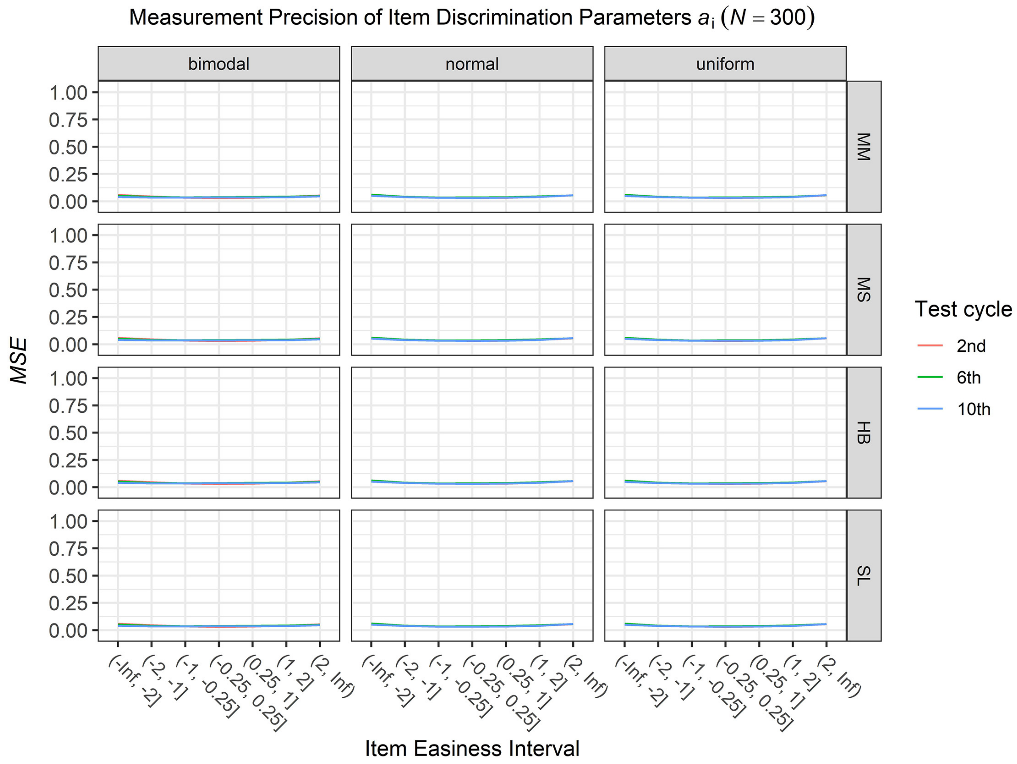


FIGURE 3 | Conditional mean squared error (*MSE*) of the item discrimination a_i for specific item easiness intervals after the 2nd, 6th, and 10th test cycle in the continuous calibration strategy with a sample size per test cycle of *N* = 300 for different common item difficulty distributions and different scale transformation methods (MM = Mean/Mean, MS = Mean/Sigma, HB = Haebara, SL = Stocking-Lord).

To answer the research questions stated above, a Monte Carlo simulation based on a full factorial design with three independent variables (IVs) was conducted. With the first IV, *difficulty distribution*, the distribution of easiness parameters d_i of the common items (normal, uniform, and bimodal with very low and very high difficulties only) was varied. The second IV, *transformation method*, compared the most common scale transformation methods (mean/mean, mean/sigma, Haebara, and Stocking-Lord) used for computing the transformation constants to conduct the scale transformation. The third IV, *sample size,* reflected the number of test takers per test cycle (N = 50; N = 100; N = 300). Because the CCS uses the responses from multiple test cycles, the number of test takers per test cycle chosen for the study is small compared to the recommendations (e.g., a minimum of 500 responses per item for the 2PL model; de Ayala, 2009). The fully crossed design comprised 3 × 4 × 3 = 36 conditions. For each of the conditions, 200 replications were conducted and analyzed with regard to various evaluation criteria (see below).

The simulations were carried out in R (R Core Team, 2018) using the "mirtCAT" package (Chalmers, 2016) for simulating adaptive tests and the "mirt" package (Chalmers, 2012) for item and person parameter estimation. Transformation constants were calculated based on the common items of consecutive test cycles using the "equateIRT" package (Battauz, 2015). The test for IPD was also conducted with the "equateIRT" package. We decided to use the "equateIRT" package in the simulations because it enables a direct import of results from the "mirt" package and offers an implemented test for IPD. The corresponding functions were called in a R script, which was written to carry out the CCS.

Simulation Procedure

Data Generation

In each replication, the discrimination parameters a_i were drawn from a lognormal distribution, $a_i \sim \log N(0,\ 0.25)$, and the easiness parameters d_i were drawn from a truncated normal distribution, $d_i \sim N(0,\ 1.5),\ \ d_i \in (-2.5,\ 2.5)$. Since this study was not designed to investigate IPD detection rates (e.g., Battauz, 2019), no IPD was simulated in the data. Therefore the true item parameters a_i and d_i remained unchanged over the test cycles.

The ability parameters of the examinees in the first test cycle in each replication were randomly drawn from a standard normal distribution, $\theta \sim N(0,\ 1)$. For the subsequent test cycles t within a replication, the ability parameters followed a normal distribution, $\theta \sim N(\mu_t,\ \sigma_t)$, whereby the mean $\mu_t \in (-0.5,\ 0.0,\ 0.5)$ and the standard deviation $\sigma_t \in (0.7,\ 1.0,\ 1.3)$ were randomly drawn. This was done to mimic the fact that examinees of different test cycles usually differ with respect to the mean and variance of their ability distribution. The examinees' responses to the items were generated in line with the 2PL model.

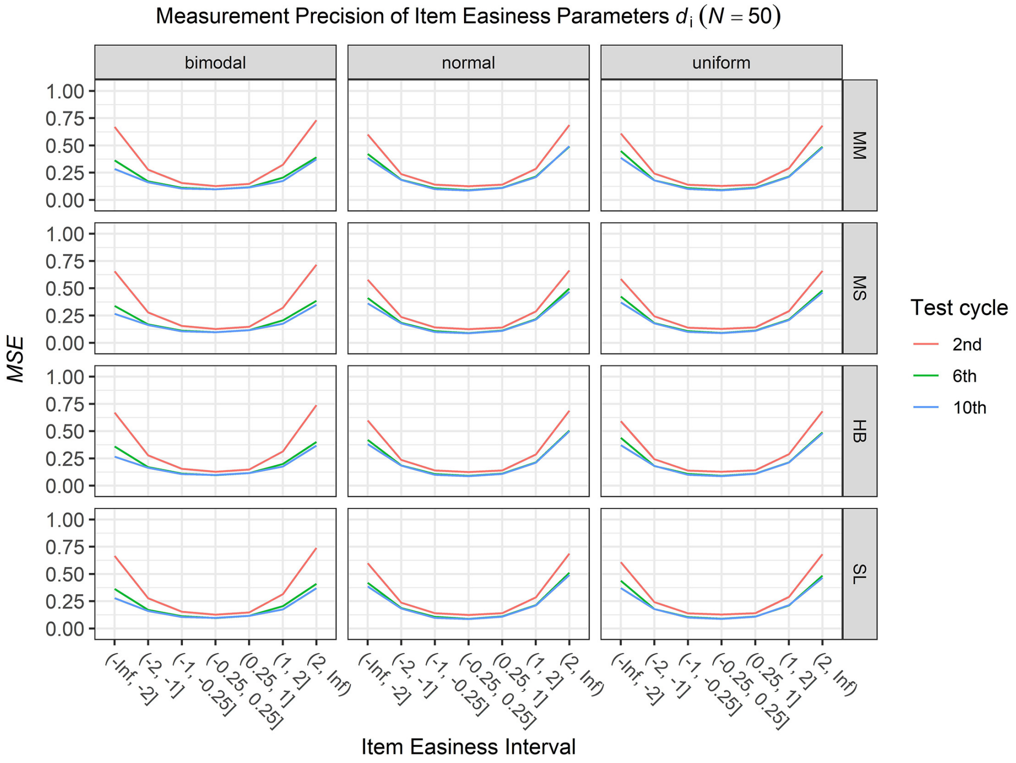


FIGURE 4 | Conditional mean squared error (*MSE*) of the item easiness d_i for specific item easiness intervals after the 2nd, 6th, and 10th test cycle in the continuous calibration strategy with a sample size per test cycle of $N = 50$ for different common item difficulty distributions and different scale transformation methods (MM = Mean/Mean, MS = Mean/Sigma, HB = Haebara, SL = Stocking-Lord).

Specification of the CCS

The CCS in the current study was applied with all seven steps proposed by Fink et al. (2018) including the IPD detection of the common items. Although no IPD was simulated in the data, in realistic settings the untested assumption of item parameter invariance is questionable. Even in the absence of IPD item parameters can significantly differ between test cycles because of sampling error. The number of test cycles within the CCS was set to 10 test cycles, whereby the first test cycle represented the initial phase and the subsequent test cycles the continuous phase. The test length was kept constant with 60 items. The calibration cluster in the continuous phase consisted of 20 items, resulting in an item pool size of $I_t = 60 + (t - 1) \cdot 20$ after the test cycle t, and a total item pool size of 240 items after the 10th test cycle. Following the recommendation of Kolen and Brennan (2014) that the number of common items should be at least 20% of the test length, the number of common items in the linking cluster was set to 15 items. Consequently, the adaptive cluster in each test cycle of the continuous phase contained 25 items. Within the adaptive cluster, the *maximum a posteriori* (MAP; Bock and Aitkin, 1981) was used as the ability estimator and the maximum information criterion (Lord, 1980) was applied for the adaptive item selection.

For the common item selection within the equating procedure, only items that had already been calibrated in the previous test cycles and that did not serve as common items in the preceding test cycle were eligible. The selection procedure for the common items differed depending on the intended distribution. For the normal distribution, the procedure of Fink et al. (2018) was applied. The eligible items were first assigned to five categories (very low, low, medium, high, and very high) based on their easiness parameters d_i. Then, five items from the “medium” category, three items each from the “low” and “high” categories, and two items from each of the extreme categories were chosen to mimic a normal distribution. For the uniform distribution, the eligible items were assigned to 15 categories based on their easiness parameters d_i and one item from each category was drawn. The interval limits of the categories were determined as quantiles of the item difficulty distribution. For the bimodal distribution, the eligible items were ordered according to their easiness parameters d_i and two subsamples were formed containing the 11 easiest and the 11 hardest items, respectively. Then, 15 items in total were randomly drawn from the two subsamples (seven easy and eight difficult items, or vice versa). As already mentioned, the selected common items in periodical assessments should be comparable also with regard to content characteristics. Content balancing approaches like the maximum priority index (Cheng and Chang, 2009) and the shadow testing approach (van der Linden and Reese, 1998) may be used for this purpose. Because no substantial impact was expected on the

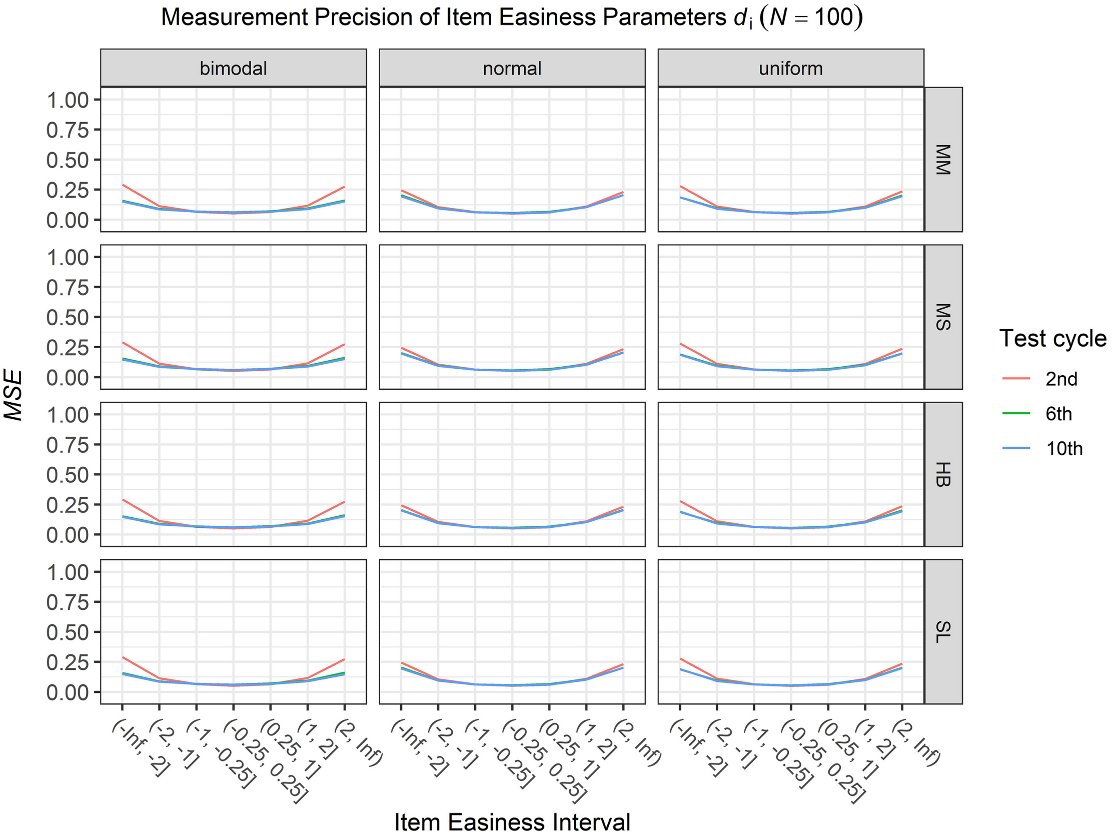


FIGURE 5 | Conditional mean squared error (*MSE*) of the item easiness d_i for specific item easiness intervals after 2nd, 6th, and 10th test cycle in the continuous calibration strategy with a sample size per test cycle of N = 100 for different common item difficulty distributions and different scale transformation methods (MM = Mean/Mean, MS = Mean/Sigma, HB = Haebara, SL = Stocking-Lord).

measurement precision of the item parameters or on the quality of the equating, content balancing was not considered as a factor in the study.

For the scale transformation, one of the four transformation methods (Mean/Mean, Mean/Sigma, Haebara, and Stocking-Lord) was applied. A modified version of Lord's chi-squared method (Lord, 1980) that is implemented in the "equateIRT" package (Battauz, 2015) was used as the test for IPD with a type I error level of 0.05. In an iterative purification process (Candell and Drasgow, 1988) of scale transformation and testing for IPD, items that showed significant IPD were removed from the set of common items. In each test cycle, MML estimation was used to obtain the item parameters for both the temporary item parameter estimation and the FCIP calibration. The lower and the upper bound for the item discrimination a_i was set to –1 and 5, respectively. For the item easiness parameters d_i, the bounds were set to –5 and 5.

Evaluation Criteria

The mean squared error (*MSE*) of the item parameters a_i and d_i, respectively, was calculated after each test cycle t as the averaged squared difference between the item parameter estimates and the true item parameters for all items I_t across all replications R = 200. Thus, a high degree of precision is denoted by low values for the *MSE*.

$$MSE_t\left(a_i\right) = \frac{1}{R^{*} I_t}\sum_{r=1}^{R}\sum_{i=1}^{I_t}\left(\hat{a}_{ir} - a_{ir}\right)^2 \tag{5}$$

$$MSE_t\left(d_i\right) = \frac{1}{R^{*} I_t}\sum_{r=1}^{R}\sum_{i=1}^{I_t}\left(\hat{d}_{ir} - d_{ir}\right)^2 \tag{6}$$

Because our aim was to evaluate whether the modified common item selection could prevent a dysfunction of the CCS in terms of more precise item parameter estimates for items with very low and very high values for d_i, the conditional *MSE* was used as a criterion. Therefore, the *MSE* was calculated for seven easiness intervals: $d_i \in \left(-Inf,\ -2\right]$, $d_i \in (-2,\ -1]$, $d_i \in (-1,\ -0.25]$, $d_i \in (-0.25,\ 0.25]$, $d_i \in (0.25,\ 1]$, $d_i \in (1,\ 2]$, and $d_i \in \left(2,\ Inf\right)$.

Three criteria were used to evaluate the equating quality. As a first criterion, we used the proportion of test cycles in which no breakdown of the common items occurred. Second, we calculated the proportion of drifted items for each of the 36 conditions. And third, we computed the accuracy (*Error*) of the scale transformation constants A and B for each replication r

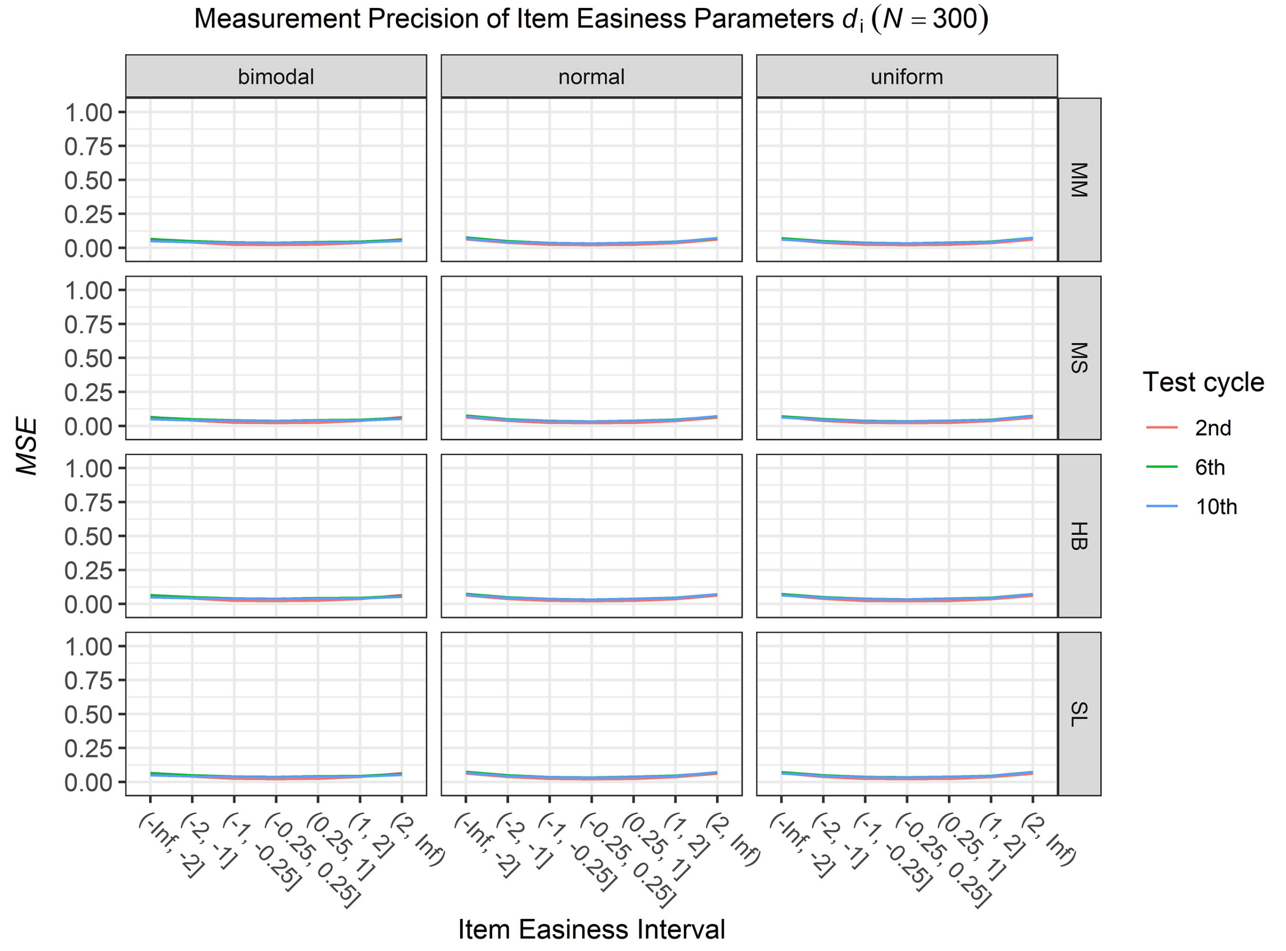


FIGURE 6 | Conditional mean squared error (*MSE*) of the item easiness d_i for specific item easiness intervals after the 2nd, 6th, and 10th test cycle in the continuous calibration strategy with a sample size per test cycle of $N = 300$ for different common item difficulty distributions and different scale transformation methods (MM = Mean/Mean, MS = Mean/Sigma, HB = Haebara, SL = Stocking-Lord).

when no breakdown occurred as the difference between the true and the estimated transformation constants for every test cycle in the continuous phase. The average of the *Error* corresponds to the Bias of the transformations constants.

$$Error\ (A_{tr}) = \left(\hat{A}_{tr} - A_{tr}\right) \quad (7)$$

$$Error\ (B_{tr}) = \left(\hat{B}_{tr} - B_{tr}\right) \quad (8)$$

The true transformation constants *A* and *B* were calculated based on the true examinees' abilities from/in all previous test cycles *p* and from/in the current test cycle *t* (Kolen and Brennan, 2014).

$$A_t = \frac{\sigma(\theta_t)}{\sigma(\theta_p)} \quad (9)$$

$$B_t = \mu(\theta_t) - A\mu(\theta_p) \quad (10)$$

The estimated transformation constants $\hat{A}_t$ and $\hat{B}_t$ were obtained based on the parameter estimates of the final set of common items from the previous and the current test cycles using one of the four scale transformation methods implemented in the "equateIRT" package (Battauz, 2015). The third criterion was calculated only for the cases where at least two common items remained after the IPD detection.

RESULTS

Note that the conditions with the mean/mean method as scale transformation method and normal distributed common items mimic the setup of the equating procedure from Fink et al. (2018).

Conditional Precision of Item Parameters

To answer the first research question regarding the precision of the item parameter estimates, we analyzed the conditional *MSE* of the item discrimination parameters a_i and the item easiness parameters d_i depending on the scale transformation method, the common item difficulty distribution, and the sample sizes per test cycle. For the sake of clarity, the results are only

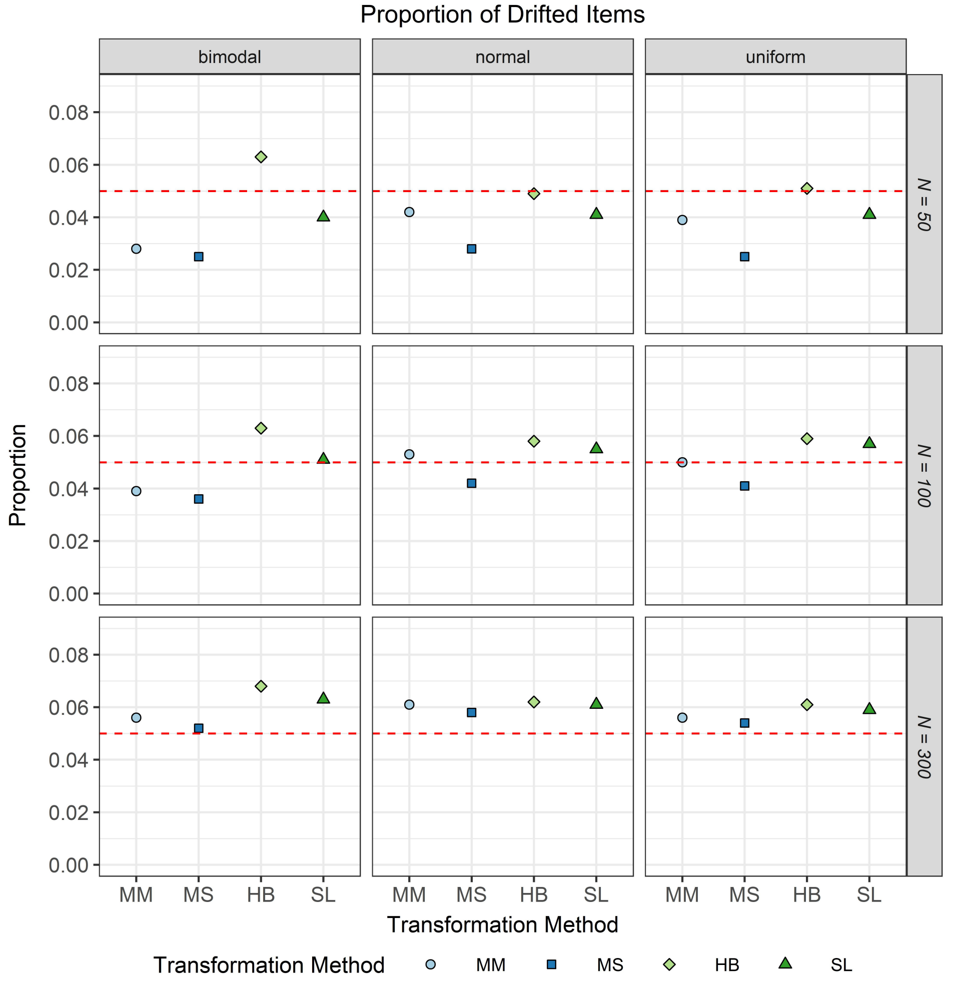


FIGURE 7 | Proportion of drifted items in the continuous calibration strategy for different sample sizes per test cycle, different common item difficulty distributions, and different scale transformation methods (MM = Mean/Mean, MS = Mean/Sigma, HB = Haebara, SL = Stocking-Lord). The dashed line represents the type I error level of 0.05.

presented for the second, the sixth, and the 10th test cycles of the CCS. **Figures 1–3** illustrate the conditional *MSE* of the item discrimination parameter estimates a_i, and **Figures 4–6** illustrate the conditional *MSE* of the item easiness parameter d_i. As can be expected based on the findings from Fink et al. (2018), the *MSE* for the item discrimination parameter estimates and the item easiness parameter estimates decreased as the number of test cycles in the CCS increased and as the sample size per test cycle increased. With regard to the precision of the item parameter estimates, no substantial differences were found between the different scale transformation methods, independent of the common item difficulty distribution and the sample size per test cycle. When a bimodal difficulty distribution of common items was chosen, the precision of the item parameter estimates for the very easy and very difficult items was higher compared to a normal or uniform difficulty distribution of common items (**Figures 1**, **4**). However, this minimal gain came at the expense of a lower precision of the item parameter estimates for items with medium difficulty. This effect was found for very small sample sizes per test cycle (N = 50), and diminished for larger sample sizes (N = 100, N = 300).

Quality of Equating

The second and third research questions focused on the equating procedure. The first evaluation criterion was the proportion of feasible equatings (at least two items remained after the IPD detection). Most striking was that over all replications for none of the test cycles a breakdown of the common items occurred. Furthermore, for all 36 conditions the median number of eligible common items over all test cycles and replications ranged from 14 to 15.

The second evaluation criterion was the proportion of drifted items. As IPD was not simulated in the study and because the type I error level of the test for IPD was set to 0.05, it was expected that approximately five percent of the common items would show significant IPD. **Figure 7** shows the proportion of drifted common items depending on the common item difficulty distribution, the scale transformation method, and the sample

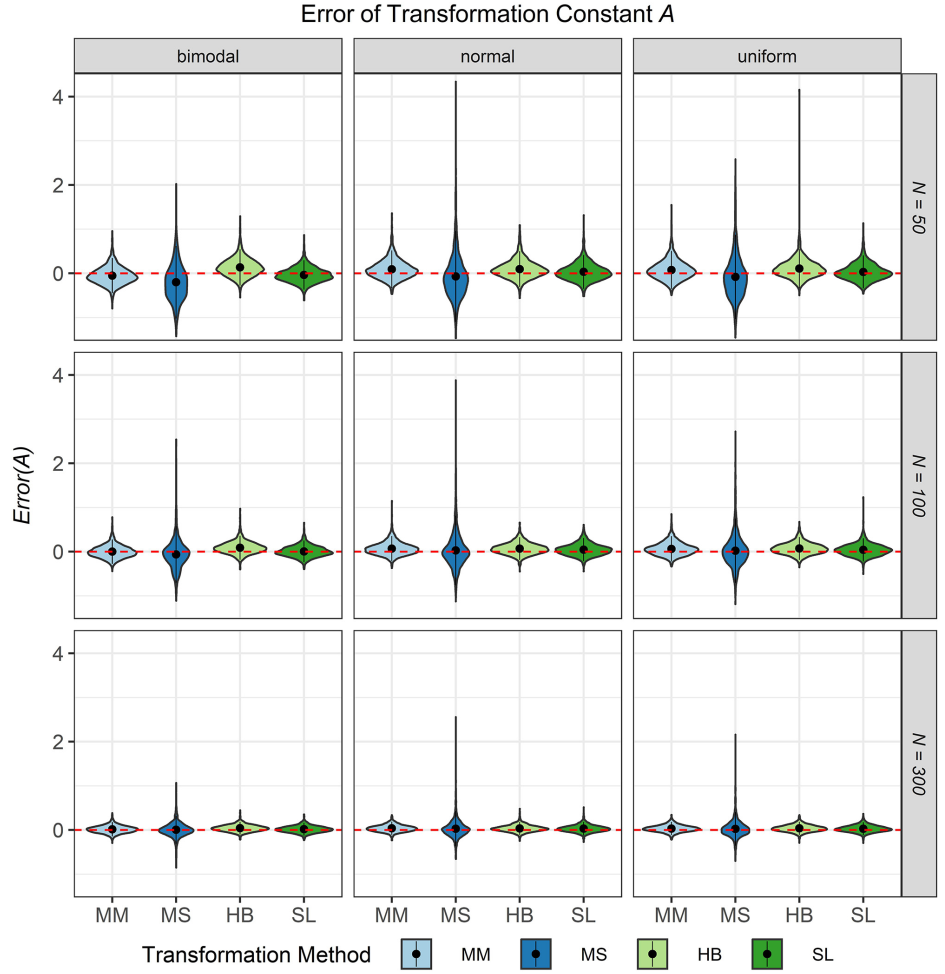


FIGURE 8 | Error of the transformation constant *A* in the continuous calibration strategy for different sample sizes per test cycle, different common item difficulty distributions, and different scale transformation methods (MM = Mean/Mean, MS = Mean/Sigma, HB = Haebara, SL = Stocking-Lord). The dot in the middle of each violin represents the bias and the width of the violin expresses the frequency of the corresponding value.

size per test cycle. It is obvious from this figure that independent of the scale transformation method and the common item difficulty distribution, the type I error rates increased with increasing sample size per test cycle. This effect was stronger for the moment/methods. Furthermore, it became apparent that if the difficulty distribution of the common items was uniform or normal, all scale transformation methods did not considerably differ from the type I error level of 0.05. The only exception to this result was the mean/sigma method which generally led to considerably smaller type I error rates when the sample size was small (N = 50). All in all, using the Stocking-Lord method resulted for all conditions in type I error rates that did not considerably differ from the type I error level of 0.05.

The third evaluation criterion was the accuracy of the transformation constants *A* and *B* when no breakdown occurred. **Figures 8**, **9** show violin plots for the *Error* of the transformation constants *A* and *B* depending on the common item difficulty distribution, the scale transformation method, and the sample size per test cycle. In violin plots, the frequency distribution of a numeric variable (e.g., bias) is expressed. Note that the average error (= *Bias*; represented by the dot in the violin) for both transformation constants *A* and *B* did not differ substantially from zero for all scale transformation methods, independent of the common item difficulty distribution and the sample size per test cycle. However, the variation of the error (represented by the height of the violin) differed between the scale transformation methods and, especially for the moment methods rather high levels of error occurred. The characteristic curve methods showed the lowest variation in error. With increasing sample size per test cycle, the variation of the error decreased, but there were still extreme levels of error for the mean/mean and the mean/sigma method.

In summary and in terms of the three research questions, the study provided the following results:

1. The difficulty distribution of the common items in the CCS did not have a substantial impact on the precision of the item parameter estimates

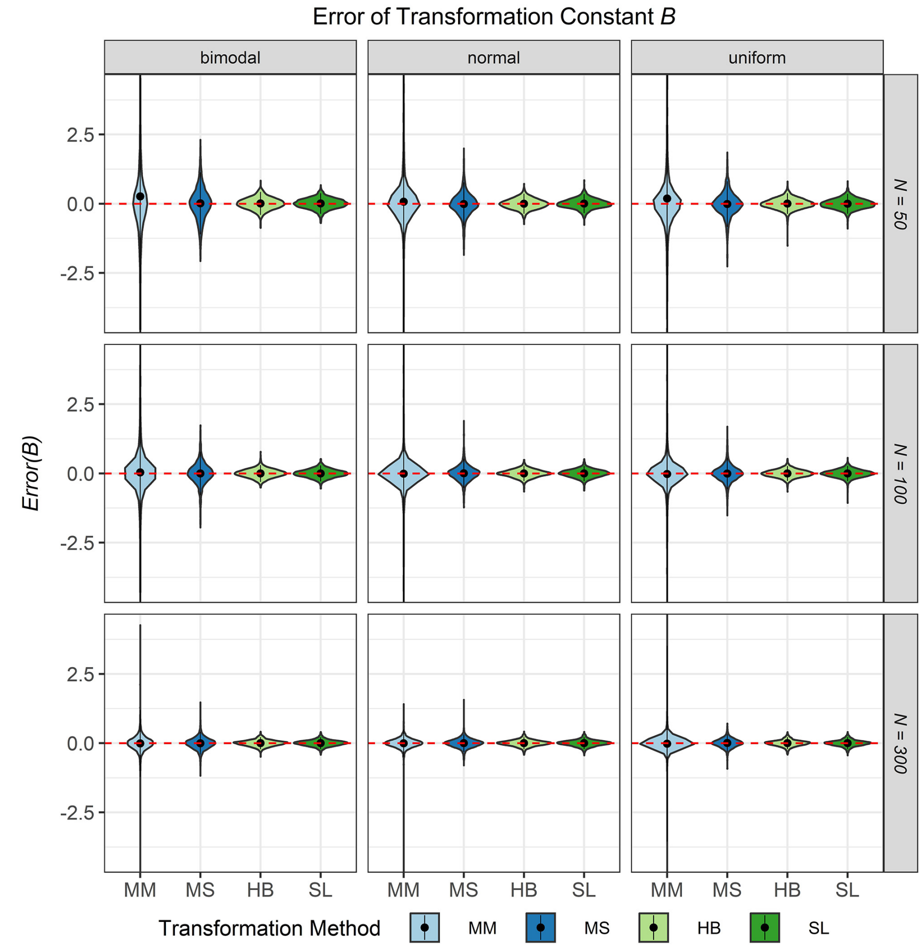


FIGURE 9 | Error of the transformation constant *B* in the continuous calibration strategy for different sample sizes per test cycle, different common item difficulty distributions, and different scale transformation methods (MM = Mean/Mean, MS = Mean/Sigma, HB = Haebara, SL = Stocking-Lord). The dot in the middle of each violin represents the bias and the width of the violin expresses the frequency of the corresponding value.

although small differences existed between the common item distributions; these differences were in opposite/varying directions for extreme and medium-ranged item easiness parameters d_i when the sample size was very small.

2. With regard to the proportion of feasible equatings (at least two common items remained after the test for IPD) no differences were found independent of the common item difficulty distributions, the scale transformation method and the sample size.
3. The characteristic curve methods outperformed the moment methods in terms of error of the transformation constant. Especially for small sample size the mean/sigma method cannot recommended.

DISCUSSION

The objective of the present study was to evaluate different setups of the equating procedure implemented in the CCS and to make/provide recommendations on how to apply these setups. For this purpose, the quality of the item parameter estimates and of the equating was examined in a Monte Carlo simulation for different common item difficulty distributions, different scale transformation methods, and different sample sizes per test cycle.

The following recommendations can be made based on the results obtained: First, no clear advantage of using any of the three common item difficulty distributions was identified. Regarding the precision of the item parameter estimates, the results show a slight increase in the precision of the item parameter estimates for items with extreme difficulties when using a bimodal common item difficulty distribution compared to a normal or uniform distribution. However, the precision of the item parameter estimates for items with medium difficulty decreased. These effects were only found for very small sample sizes per test cycle (N = 50) and no differences were found for larger sample sizes (N = 100, N = 300). Furthermore, the use of different scale transformation methods did not have a substantial effect on the precision of the item parameter estimates.

Note that exposure control methods (e.g., Sympson and Hetter, 1985; Revuelta and Ponsoda, 1998; Stocking and Lewis, 1998) might be an alternative to increase the number of responses to items with extreme difficulty levels and, in consequence, the precision of the item parameter estimates for these items. However, using these methods would sacrifice adaptivity to a certain degree and, thus, the efficiency of the computerized adaptive test (e.g., Revuelta and Ponsoda, 1998). This is even more relevant to tests assembled within the partly adaptive CCS, because only one of the three cluster types used is based on an adaptive item selection. Furthermore, in the early stages of the CCS, the item pool is rather small, which also limits the adaptivity of the tests. For these reasons, it can be expected that exposure control methods do not offer an ideal option for the CCS to increase the precision of item parameter estimates for items with extreme difficulties. This point might be examined by future research.

Second, with respect to the quality of the equating, no difference was found for the scale transformation methods with regard to the proportion of feasible equatings independent of the common item difficulty distribution used and the sample size available per test cycle. The rule for evaluating an equating as feasible (at least two common items remained after the test for IPD) is worthy of discussion because of two reasons: first, with a small number of remaining common items, the equating procedure is more prone to sampling error (Wingersky and Lord, 1984) and second, it is rather unlikely that the content of the item pool is adequately reflected by the remaining common items. However, even if the criterion for evaluating an equating as feasible had been set to ten remaining common items, the proportion of feasible equatings would be at least 99% in all conditions. With regard to the type I error rate and the error of the transformation constant the characteristic curve methods outperformed the moment methods especially for small sample sizes. This is in line with the result of Ogasawara (2002) who found that the characteristic curve methods are less affected by imprecise item parameter estimates and lead to more accurate transformation than moment methods. Among the characteristic curve methods the Stocking-Lord method was slightly better than the Haebara method in almost all conditions. Thus, although our results do not facilitate a clear recommendation regarding the most favorable common item difficulty distribution, they do enable a clear recommendation in terms of the preferred scale transformation method: The Stocking-Lord method should be used as the scale transformation method within the CCS.

AUTHOR CONTRIBUTIONS

SB conceived the study, conducted the statistical analyses, drafted the manuscript, and approved the submitted version. AFi performed substantial contribution to the conception of the study, contributed to the programming needed for the simulation study (R), reviewed the manuscript critically for important intellectual content, and approved the submitted version. CS performed substantial contributions to the interpretation of the study results, reviewed the manuscript critically for important intellectual content, and approved the submitted version. AFr provided advise in the planning phase of the study, reviewed the manuscript critically for important intellectual content, and approved the submitted version.

FUNDING

The research reported in the article was supported by a grant from the German Federal Ministry of Education and Research (Ref: 16DHL1005).

REFERENCES

Baker, F. B., and Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *J. Educ. Meas.* 28, 147–162. doi: 10.1111/j.1745-3984.1991.tb00350.x

Battauz, M. (2015). equateIRT: an R package for IRT test equating. *J. Stat. Softw.* 68, 1–22. doi: 10.18637/jss.v068.i07

Battauz, M. (2018). "Simultaneous equating of multiple forms," in *Quantitative Psychology*, eds M. Wiberg, S. Culpepper, R. Janssen, J. González, and D. Molenaar (Cham: Springer), 121–130.

Battauz, M. (2019). On wald tests for differential item functioning detection. *Stat. Methods Appl.* 28, 121–130. doi: 10.1007/s10260-018-00442-w

Birnbaum, A. (1968). "Some latent trait models and their use in inferring an examinee's ability," in *Statistical Theories of Mental Test Scores*, eds F. M. Lord and M. R. Novick (Reading, MA: Addison-Wesley), 395–479.

Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika* 46, 443–459. doi: 10.1007/BF02293801

Born, S., and Frey, A. (2017). Heuristic constraint management methods in multidimensional adaptive testing. *Educ. Psychol. Meas.* 77, 241–262. doi: 10.1177/0013164416643744

Candell, G. L., and Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Appl. Psychol. Meas.* 12, 253–260. doi: 10.1177/014662168801200304

Chalmers, R. P. (2012). mirt: a multidimensional item response theory package for the R environment. *J. Stat. Softw.* 48, 1–29. doi: 10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *J. Stat. Softw.* 71, 1–39. doi: 10.18637/jss.v071.i05

Cheng, Y., and Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *Br. J. Math. Stat. Psychol.* 62, 369–383. doi: 10.1348/000711008X304376

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York, NL: Guilford.

Fink, A., Born, S., Spoden, C., and Frey, A. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychol. Test Assess. Model.* 60, 327–346.

Goldstein, H. (1983). Measuring changes in educational attainment over time: problems and possibilities. *J. Educ. Meas.* 20, 369–377. doi: 10.1111/j.1745-3984.1983.tb00214.x

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Jpn. Psychol. Res.* 22, 144–149. doi: 10.4992/psycholres1954.22.144

Hanson, B. A., and Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Appl. Psychol. Meas.* 26, 3–24. doi: 10.1177/0146621602026001001

He, W., and Reckase, M. D. (2014). Item pool design for an operational variable-length computerized adaptive test. *Educ. Psychol. Meas.* 74, 473–494. doi: 10.1177/0013164413509629

Hu, H., Rogers, W. T., and Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Appl. Psychol. Meas.* 32, 311–333. doi: 10.1177/0146621606292215

Kaskowitz, G. S., and de Ayala, R. J. (2001). The effect of error in item parameter estimates on the test response function method of linking. *Appl. Psychol. Meas.* 25, 39–52. doi: 10.1177/01466216010251003

Kim, S. H., and Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *J. Educ. Meas.* 29, 51–66. doi: 10.1111/j.1745-3984.1992.tb00367.x

Kolen, M. J., and Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices*, 3rd Edn. New York, NY: Springer, doi: 10.1007/978-1-4939-0317-7_10

Lord, F. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Loyd, B. H., and Hoover, H. D. (1980). Vertical equating using the rasch model. *J. Educ. Meas.* 17, 179–193. doi: 10.1111/j.1745-3984.1980.tb00825.x

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *J. Educ. Meas.* 14, 139–160. doi: 10.1111/j.1745-3984.1977.tb00033.x

Miller, G. E., and Fitzpatrick, S. J. (2009). Expected equating error resulting from incorrect handling of item parameter drift among the common items. *Educ. Psychol. Meas.* 69, 357–368. doi: 10.1177/0013164408322033

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Econ. Rev.* 51, 1–23.

Ogasawara, H. (2002). Stable response functions with unstable item parameter estimates. *Appl. Psychol. Meas.* 26, 239–254. doi: 10.1177/0146621602026003001

R Core Team (2018). *R: A Language and Environment for Statistical Computing [Software]*. Vienna: R Foundation for Statistical Computing.

Revuelta, J., and Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *J. Educ. Meas.* 35, 311–327. doi: 10.1111/j.1745-3984.1998.tb00541.x

Segall, D. O. (2005). "Computerized adaptive testing," in *Encyclopedia of Social Measurement*, ed. K. Kempf-Leonard (Boston: Elsevier Academic), 429–438. doi: 10.1016/b0-12-369398-5/00444-8

Stocking, M. L., and Lewis, C. L. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *J. Educ. Behav. Stat.* 23, 57–75. doi: 10.3102/10769986023001057

Stocking, M. L., and Lord, F. M. (1983). Developing a common metric in item response theory. *Appl. Psychol. Meas.* 7, 201–210. doi: 10.1177/014662168300700208

Sympson, J. B., and Hetter, R. D. (1985). "Controlling item exposure rates in computerized adaptive testing," in *Proceedings of the 27th Annual Meeting of the Military Testing Association*, (San Diego, CA: Navy Personnel Research and Development Center), 973–977.

Thissen, D., Steinberg, L., and Wainer, H. (1988). "Use of item response theory in the study of group difference in trace lines," in *Test Validity*, eds H. Wainer and H. Braun (Hillsdale, NJ: Lawrence Erlbaum Associates).

Thompson, N. A., and Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Pract. Assess. Res. Eval.* 16:9.

Vale, C. D., Maurelli, V. A., Gialluca, K. A., Weiss, D. J., and Ree, M. J. (1981). *Methods for Linking Item Parameters (AFHRL-TR-81-10)*. Brooks Air Force Base TX: Air Force Human Resources Laboratory.

van der Linden, W. J. (2016). *Handbook of Item Response Theory*, Vol. 1. London: Chapman and Hall.

van der Linden, W. J., and Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Appl. Psychol. Meas.* 22, 259–270. doi: 10.1177/01466216980223006

Weeks, J. P. (2010). plink: an r package for linking mixed-format tests using IRT-based methods. *J. Stat. Softw.* 35, 1–33. doi: 10.18637/jss.v035.i12

Wingersky, M. S., and Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Appl. Psychol. Meas.* 8, 347–364. doi: 10.1177/014662168400800312

Yousfi, S., and Böhme, H. F. (2012). Principles and procedures of considering item sequence effects in the development of calibrated item pools: conceptual analysis and empirical illustration. *Psychol. Test Assess. Model.* 54, 366–393.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling Editor declared a shared affiliation, though no other collaboration, with one of the authors AFr at the time of review.

frontiers in Psychology

ORIGINAL RESEARCH
published: 07 June 2019
doi: 10.3389/fpsyg.2019.01280

Collaborative Problem Solving: Processing Actions, Time, and Performance

Paul De Boeck[1,2]* *and Kathleen Scalise*[3]*

[1] Department of Psychology, The Ohio State University, Columbus, OH, United States, [2] Department of Psychology, KU Leuven, Leuven, Belgium, [3] Department of Educational Methodology, Policy, and Leadership, University of Oregon, Eugene, OR, United States

This study is based on one collaborative problem solving task from an international assessment: the Xandar task. It was developed and delivered by the Organization for Economic Co-operation and Development Program for International Student Assessment (OECD PISA) 2015. We have investigated the relationship of problem solving performance with invested time and number of actions in collaborative episodes for the four parts of the Xandar task. The parts require the respondent to collaboratively plan a process for problem solving, implement the process, reach a solution, and evaluate the solution (For a full description, see the Materials and Methods section, "Parts of the Xandar Task.") Examples of an action include posting to a chat log, accessing a shared resource, or conducting a search on a map tool. Actions taken in each part of the task were identified by PISA and recorded in the data set numerically. A confirmatory factor analysis (CFA) model looks at two types of relationship: at the level of latent variables (the factors) and at extra dependencies, which here are direct effects and correlated residuals (independent of the factors). The model, which is well-fitting, has three latent variables: actions (A), times (T), and level of performance (P). Evidence for the uni-dimensionality of performance level is also found in a separate analysis of the binary items. On the whole for the entire task, participants with more activities are less successful and faster, based on the United States data set employed in the analysis. By contrast, successful participants take more time. By task part, the model also investigates relationships between activities, time, and performance level within the parts. This was done because one can expect dependencies within parts of such a complex task. Results indicate some general and some specific relationships within the parts, see the full manuscript for more detail. We conclude with a discussion of what the investigated relationships may reveal. We also describe why such investigations may be important to consider when preparing students for improved skills in collaborative problem solving, considered a key aspect of successful 21st century skills in the workplace and in everyday life in many countries.

Keywords: problem solving, strategy, factor model, measurement, collaboration

OPEN ACCESS

Edited by:
Ronny Scherer,
University of Oslo, Norway

Reviewed by:
Sharlene D. Newman,
Indiana University Bloomington,
United States
Maria Bolsinova,
University of Amsterdam, Netherlands

Correspondence:
Paul De Boeck
deboeck.2@osu.edu
Kathleen Scalise
kscalise@uoregon.edu

Specialty section:
This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Psychology

Received: *18 December 2018*
Accepted: *15 May 2019*
Published: *07 June 2019*

Citation:
De Boeck P and Scalise K (2019)
Collaborative Problem Solving:
Processing Actions, Time,
and Performance.
Front. Psychol. 10:1280.
doi: 10.3389/fpsyg.2019.01280

INTRODUCTION

The construct explored here, collaborative problem solving (CPS), was first introduced to the Program for International Student Assessment (PISA) in 2015. Attempts to explore process data collected in complex activities such as CPS are emerging rapidly in education. Yet which models might best fit process data and the analytic techniques to employ to investigate patterns in the data are not well understood at this time. So here we investigate whether relationships seen in the actions taken by PISA respondents, as coded by PISA, might shed light on approaches for modeling complex CPS tasks.

In the CPS task released by PISA, the Xandar task, there are four parts. The parts of the task require the respondent to collaborate to plan a process for problem solving, implement the process, reach a solution, and evaluate the solution. (For a full description of these parts, see the Materials and Methods section, "Parts of the Xandar Task.") Examples of actions in Part 1, for instance, include posting to a chat log, accessing a shared resource, or conducting a search on a shared map tool.

In each of the parts, process data are available on time spent and number of actions, as well as on the performance on specific items within the four parts. We explore modeling these Xandar data to address three research questions:

> RQ1. Does a factor model employing process data (actions and time) support evidence for a latent variable differentiation between the types of process data (actions, time) and between the latter two and quality of performance? The expected latent variables are Actions, Time, and Performance.
>
> RQ2. Do extra dependencies at the level of the observed variables improve model fit, including direct effects and correlated residuals (independent of the factors)? If they do, they reveal direct relationships between process aspects and performance, independent of the latent variables. These direct relationships are indications of the dynamics underlying collaborative problem solving, whereas the latent variables and their correlations inform us about global individual differences in process approaches and performance.
>
> RQ3. Can the performance also be considered as uni-dimensional at the specific level of the individual items (from all four Xandar parts)?

In this Xandar investigation, each factor (latent variable) is composed of four corresponding measures from the four Xandar parts. Data are fit with a latent variable model to answer RQ1. Dependencies within parts can be expected between the three measures. So we address the extra dependencies in RQ2. The dependencies are not only considered for methodological reasons when variables stem from the same part, but they may also reveal how subjects work on the tasks. Finally, because a good-fitting factor model would imply uni-dimensionality of the performance sum scores from the four parts, we also explore uni-dimensionality at the level of the individual items in RQ3.

Sections in this paper first discuss the PISA efforts to explore problem solving in 2012 and 2015 assessments, then offer a brief summary of the literature on CPS. Next in the Materials and Methods section, we discuss the PISA 2015 collaborative complex problem solving released task, "Xandar," including the availability of the released code dictionary and data set. In the Results and Discussion, we model United States data from the Xandar task and report results to address the three research questions.

PISA AND A BRIEF SUMMARY OF LITERATURE ON CPS

The PISA 2015 CPS construct, which included measuring groups in collaboration, was built on PISA's 2012 conception of individual problem solving (OECD, 2014). In PISA 2012, some student individual characteristics related to individual problem solving were measured. These measures were openness to learning, perseverance, and problem solving strategies.

For the 2015 PISA collaborative framework (OECD, 2013), the construct of problem solving was extended from 2012 in order to include measures of group collaboration. For this new assessment in 2015, it was recognized that the ability of an individual to be successful in many modern situations involves participating in a group. Collaboration was intended to include such challenges as communicating within the group, managing conflict, organizing a group, and building consensus, as well as managing progress on a successful solution.

The PISA framework described the importance of improving collaboration skills for students (Rummel and Spada, 2005; Vogel et al., 2016) The measurement of collaboration skills was at the heart of problem solving competencies in the PISA CPS 2015 framework. The framework specified first that the competency being described remained the capacity of an individual, not the group. Secondly, the respondent must effectively engage in a process whereby two or more agents attempt to solve a problem, where the agents can be people or simulations. Finally, the collaborators had to show efficacy by sharing the understanding and effort required to come to a solution, such as pooling knowledge to reach solutions.

Approaches to gathering assessment evidence cited by the PISACPS framework (OECD, 2013) ranged from allowing *actions* during collaboration to evaluating the *results* from collaboration. Measures of collaboration in the research literature include solution success, as well as processes during the collaboration (Avouris et al., 2003). *In situ* observables for such assessments could include analyses of log files in which the computer keeps a record of student activities, sets of intermediate results, and paths taken along the way (Adejumo et al., 2008). Group interactions also offer relevant information (O'Neil et al., 1997), including quality and type of communication (Cooke et al., 2003; Foltz and Martin, 2008; Graesser et al., 2008) and judgments (McDaniel et al., 2001).

The international Assessment and Teaching for twenty-first century Skills (ATC21S) project also examined the literature on disposition to collaboration and to problem solving in online environments. ATC21S described how interface design feature issues and the evaluation of CPS processes interact in the online collaboration setting (Scalise and Binkley, 2009; Binkley et al., 2010, 2012).

In the PISA 2015 CPS assessment, a student's collaborative problem-solving ability is assessed in scenarios where the student must solve a problem. For collaboration, the problem is solving working with "agents," or computer avatars that simulate collaboration. The CPS framework describes that a problem need not be subject-matter specific task,. Rather it could also be as a partial task in an everyday problem. Examples of subject-matter specific problem solving include setting up a sustainable fish farm in science, planning the construction of a bridge using engineering and mathematics, or writing a persuasive letter using language arts and literacy Examples of an "everyday" problem include communicating with others to delegate roles during collaboration for event planning, monitoring to ensure a group remains on task, and evaluating whether collaboration is complete. All these actions can be directed toward the ultimate goal.

In the PISA 2015 perspective, assessment is continuous throughout the unit and can incorporate student's interactions with the digital agents. Each student response on a traditional question follows a stream of actions during which the student has chosen how to interact and collaborate with standardized agents in each particular task situation. Very few of the collaborative actions and tasks are released by PISA, but the *number* of collaborative actions in each part of the task are released and made available in the PISA data sets. So here we accept that PISA has coded the action as taking place, and analyze the numeric results provided.

MATERIALS AND METHODS

Parts of the Xandar Task

Here we analyze numeric data provided for the PISA 2015 Xandar unit (OECD, 2017a,b). In the unit Xandar:

> "A three-person team consisting of the student test-taker and two computer agents takes part in a contest where [the team] must answer questions about the fictional country of Xandar. The questions [involve] Xandar's geography, people and economy. This unit involves decision-making and coordination tasks, requires consensus-building collaboration, and has an in-school, private, and non-technology-based context."

Xandar is a fictional planet appearing in comic books published by Marvel Comics. In the PISA Xandar task, it is treated as a mythical location to be investigated collaboratively. The Xandar task has four parts:

- Part 1 – Agreeing on a Strategy. This part of the Xandar activity familiarizes the student with how the contest will proceed, the chat interface and the task space including buttons that students can click to take actions in particular situations and a scorecard that monitors team progress. In Part 1, the student is assigned to work in a team with digital agents named Alice and Zach. A variety of actions are available. The respondent and the agents interact to generate a stream of actions. The respondent is expected to follow the rules of engagement provided for the contest and to effectively establish collaborative and problem-solving strategies that were the goal of Part 1.
- Part 2 – Reaching a Consensus Regarding Preferences. In this part of the Xandar activity, group members should take responsibility for the contest questions in one subject area (Xandar's geography, people, or economy). The team members must apportion the subject areas among themselves. The agents begin by disagreeing. The student has opportunities to help resolve the disagreement, can take a variety of actions, and the goal is to establish common understanding.
- Part 3 – Playing the Game Effectively. In this part of the Xandar activity, group members begin playing the game by answering geography contest questions together. The group has the opportunity to choose among answers, during which the agents interject questions, pose concerns and even violate game rules. The student exhibits collaborative problem solving strategies through actions and responses.
- Part 4 – Assessing Progress. In this part of the Xandar activity, agent Alice has posed a question about its progress. The student responds with an evaluation. Regardless of the student's answer, agent Zach indicates he is experiencing trouble information foraging for his assigned subject area, economy. Responses and actions take place regarding both evaluating and supporting group members.

Each of the four parts comes with a number of items to score the performance. The complete Xandar released task is presented in an OECD PISA report that illustrates the items that students faced in the 2015 PISA collaborative problem-solving assessment (OECD, 2016). The released code dictionary and data are also available on the 2015 PISA website. We do not repeat the Xandar information here (due in part to copyright), but summarize only. The Xandar released unit presents:

- a screenshot of each item
- the correct action(s) or response to the item
- an explanation as to why the action or response is correct
- the skills that are examined by the item
- alignments describing the difficulty of the item.

Sample

As described earlier, this study employed data publicly released from the Organization for Economic Co-operation and Development Program for International Student Assessment (OECD PISA) for the optional collaborative problem solving (CPS) assessment. It was administered in 2015 to nationally representative samples of approximately age 15 students. Since PISA is designed to have systematically missing data in a matrix sample, only students who took the Xandar task were included. Students were sampled according to the PISA sample frame. Data analyzed here are representatively sampled United States participants from the Xandar released task. See **Table 1** for descriptives by age, gender and race/ethnicity of the United States Xandar task sample used.

From the 994 students who took the Xandar task, 986 have complete Xandar data. The descriptive statistics and all analyses are based on N = 986. (Note that limitations to be discussed later in this manuscript include only United States data examined to date in this exploration. Extensions to more countries and comparisons across countries are an exciting and interesting potential to the work. However, the international extensions are out of scope for this article.) For the purposes of the current study, the school variable was not employed. All students were treated as one group.

Regarding ethical approval and consent for human subjects data collection in PISA, OECD gains ethical approval and consent through PISA processes. Processes are established in coordination with each country for public release of some de-identified data collected in PISA main study assessments. Data sets made available for release are intended for purposes of secondary research. The CPS data set used here is available through the OECD data repository website[1].

As discussed earlier, for the Xandar task, released data are available for actions, time and level of performance. The data for the current study included four indicators each of CPS actions taken (parts 1–4), time taken (parts 1–4), and success scores (parts 1–4). These become the three latent traits, or factors, in this study. To measure CPS actions, we used number of collaboration actions as measured by the data provided in the log transformation of C1A, C2A, C3A, and C4A. "C" indicates this was a collaborative assessment, the numeral indicates the Xandar part, and "A" indicates number of actions taken. To measure timing, we used timing as measured by data provided in the log transformation of C1T, C2T, C3T, and C4T. "C" indicates this was a collaborative assessment, the numeral indicates the Xandar part, and T indicates time taken. To measure student success, we used the sum of the binary item response success scores for each of the four parts, C1P, C2P, C3P, and CP4 (based on 5, 3, 2, and 2 items within the Xandar parts).

Exploratory data analysis following log transformation as described above for some variables revealed only minor deviations from normality. Skewness between −2 to 2 was used for all observed variables (Cohen et al., 2002). Note, however, that this is not a strongly conservative range, as discussed in the limitations. So we also report for this study skewness with all observed variables approximately in the range −1 to 1 except for C1A (1.52) and C2A (1.48). Due to no major levels of deviation, the analysis proceeded without further transformation to the observed variables. Other descriptives for all observed variables are provided in **Table 2**.

We fit the model using lavaan (Rosseel, 2012) in R version 3.5.1 (R Core Team, 2018). We used the weighted least squares "WLSMV" option which employs the diagonally weighted least squares (DWLS) estimator with robust standard errors and a mean and variance adjusted test statistic. We have estimated a confirmatory factor analysis (CFA) model with three factors (each with standardized latent variables). The factors are Actions, Time, and Performance. Each one has the four corresponding measures from the four Xandar parts.

Because dependencies within parts can be expected between the three measures, some parameters were added to the model. They are direct within-part effects of actions on time (more actions implies more time), direct within-part effects of performance on time (better performance may take more time), and correlated residuals for actions and performance within each part (exploring the relationship between actions and performance level).

Direct effects and residual correlations are two different types of dependencies. Direct effects are effects of one variable on another (e.g., of Y_1 on Y_2). The two directions, $Y_1 \rightarrow Y_2$ and $Y_2 \rightarrow Y_1$, are not mathematically equivalent. Correlated residuals are equivalent with the effect of a residual of one variable on the other variable (e.g., of ε_{Y1} on Y_2). the two directions are mathematically equivalent and equivalent with the covariance of the residuals. To be clear, neither of the dependencies prove a causality relation. A causal hypothesis

TABLE 1 | Descriptives for collaborative problem solving Xandar assessment for the United States sample.

Descriptive	N	Percentage
Total sample	986	100%
Birth year		
1999	479	48.58%
2000	498	50.51%
Missing	9	<1%
Gender (binary only in PISAB)		
Male	503	51.01%
Female	474	48.07%
Missing	9	<1%
Race/Ethnicity		
White, not Hispanic	409	41.48%
Black or African American	138	14.00%
Hispanic or Latino	314	31.85%
Asian	36	3.65%
Multi-racial	67	6.80%
Other	7	<1%
Missing	15	1.52%

TABLE 2 | Descriptives for observed variables.

Variable	Mean	SD	Min	Max	% Missing
C1T	11.70	0.30	10.85	12.87	0.00
CIA	2.58	0.33	0.00	5.12	0.00
C2T	11.20	0.29	9.52	13.17	0.00
C2A	2.14	0.28	0.00	3.74	0.00
C3T	11.19	0.31	9.71	12.17	0.00
C3A	2.76	0.39	0.00	4.03	0.00
C4T	10.19	0.45	8.78	11.51	0.00
C4A	1.61	0.27	0.69	3.58	0.00
C1P	3.49	1.35	0	5	0.00
C2P	1.95	0.86	0	3	0.00
C3P	1.03	0.56	0	2	0.00
C4P	0.99	0.74	0	2	0.00

[1] www.oecd.org/pisa/data/

can be at the basis of hypothesizing a direct effect, whereas correlated residuals can be used for explorative purposes, without specifying a direction. For the present study, we hypothesized that more actions take more time and that a higher level of performance requires more time. For number of actions and level of performance we explore the dependency with correlated residuals.

See the row heads of **Tables 3**, **4** and **Figure 1** for a definition of the model estimated. It includes the latent variable structure as well as the dependencies. The model can also be derived from the R code for the analysis, which is available in the **Supplementary Material**.

RESULTS

In this section we describe the results of the modeling. With the dependencies as described in the Methods section added to the model, the model fit was good (close), with a TLI of 0.95 and RMSEA of 0.038 (90% CI 0.029 to 0.048). Without the dependencies (without the eight direct effects and four residual correlations), the model fit is clearly worse, with a TLI of 0.574 and RMSEA of 0.112 (90% CI 0.104 to 0.119). These results address RQ1 and RQ2.

TABLE 3 | CFA factor loadings Xandar measures.

Variable	Estimate	SE	*z*	*p*	Standardized
Action factor					
CIA	0.19	0.02	7.84	< 0.001	0.57
C2A	0.17	0.02	7.61	< 0.001	0.63
C3A	0.22	0.03	7.91	< 0.001	0.58
C4A	0.05	0.01	4.09	< 0.001	0.19
Time factor					
C1T	0.26	0.01	20.48	< 0.001	0.85
C2T	0.24	0.01	17.71	< 0.001	0.84
C3T	0.17	0.01	13.20	< 0.001	0.54
C4T	0.19	0.02	11.87	< 0.01	0.43
Performance factor					
C1P	0.87	0.05	16.81	< 0.01	0.64
C2P	0.47	0.03	14.09	< 0.01	0.55
C3P	0.25	0.02	10.58	< 0.01	0.44
C4P	0.28	0.03	10.39	< 0.01	0.38

TABLE 4 | Extra dependencies in CFA model for Xandar measures.

Variables	Estimate	SE	*z*	*p*	Standardized
CIA→CIT	0.41	0.047	8.57	< 0.001	0.45
C2A→C2T	0.52	0.084	6.11	< 0.001	0.49
C3A→C3T	0.30	0.046	6.56	< 0.001	0.37
C4A→C4T	0.49	0.061	7.94	< 0.001	0.29
CIP→CIT	0.01	0.01	1.10	> 0.05	0.04
C2P→C2T	0.00	0.03	–0.21	> 0.05	–0.01
C3P→C3T	0.00	0.02	–0.43	> 0.05	–0.01
C4P→C4T	0.26	0.02	15.28	< 0.001	0.42
CIP→CIA	–0.04	0.01	–3.14	< 0.01	–0.16
C2P↔C2A	0.01	0.81	0.42	> 0.05	0.04
C3P↔C3A	0.01	0.88	0.38	> 0.05	0.04
C4P↔C4A	0.06	0.01	9.67	< 0.001	0.32

→indicates direct effects and ↔ indicates correlated residuals.

The correlations between the latent variables are −0.473, $p<0.001$ (Actions and Time), −0.732, $p < 0.001$ (Actions and Performance), and 0.190, $p < 0.01$ (Time and Performance). The loadings and dependencies are shown in **Tables 3**, **4**, respectively. As expected, the indicators of actions, time, and performance all showed significant positive factor loadings on the corresponding factors (see **Table 3**). The standardized coefficients in the last column indicate that the loadings of the Part 4 indicators are lower than those of the other three parts: 0.19 (Actions), 0.43 (Time), and 0.38 (Performance).

Table 4 shows the estimates of the dependencies:

- Number of activities makes time longer: a significant positive effect was found for all four parts.
- A significant positive effect of performance on time was found only for Part 4. For the other parts the effect was almost zero.
- Number of activities and performance levels have significant correlated residuals for two parts. For explorative reasons the dependencies were not tested with a direction but with correlated residuals instead. The results were found to be different depending on the part. Results showed negative dependency for Part 1, a positive dependency for Part 4, and an almost zero dependency for the Parts 2 and 3.

Although the factor model with these dependencies fits well, we wanted to check whether the performance is also uni-dimensional at the level of the individual items (RQ3). Uni-dimensionality of the four sum scores as implied by the factor model, does not imply uni-dimensionality at the level of the 12 individual binary items. This is especially because the items represent four processes (exploring and understanding, representing and formulating, planning and executing, and monitoring and reflecting) and three competencies (establishing and maintaining shared understanding, taking appropriate action to solve the problem, and establishing and maintaining team organization), but not with a perfectly crossed design.

The answer to the dimensionality question based on the analysis with this data set is that the 12 items can be considered as uni-dimensional based on the empirical data, although they are designed to tap on a diversity of processes and competencies. The uni-dimensional model fit was good (close), with a TLI of 0.94 and RMSEA of 0.037 (90% CI 0.029, 0.046). The uni-dimensional model is the result of an ordinal confirmatory factor model for the binary items using WLSMV and the same lavaan version as for the earlier analysis. For the delta parameterization the loadings vary between 0.272 and 0.776 and they are all significant ($p < 0.001$).

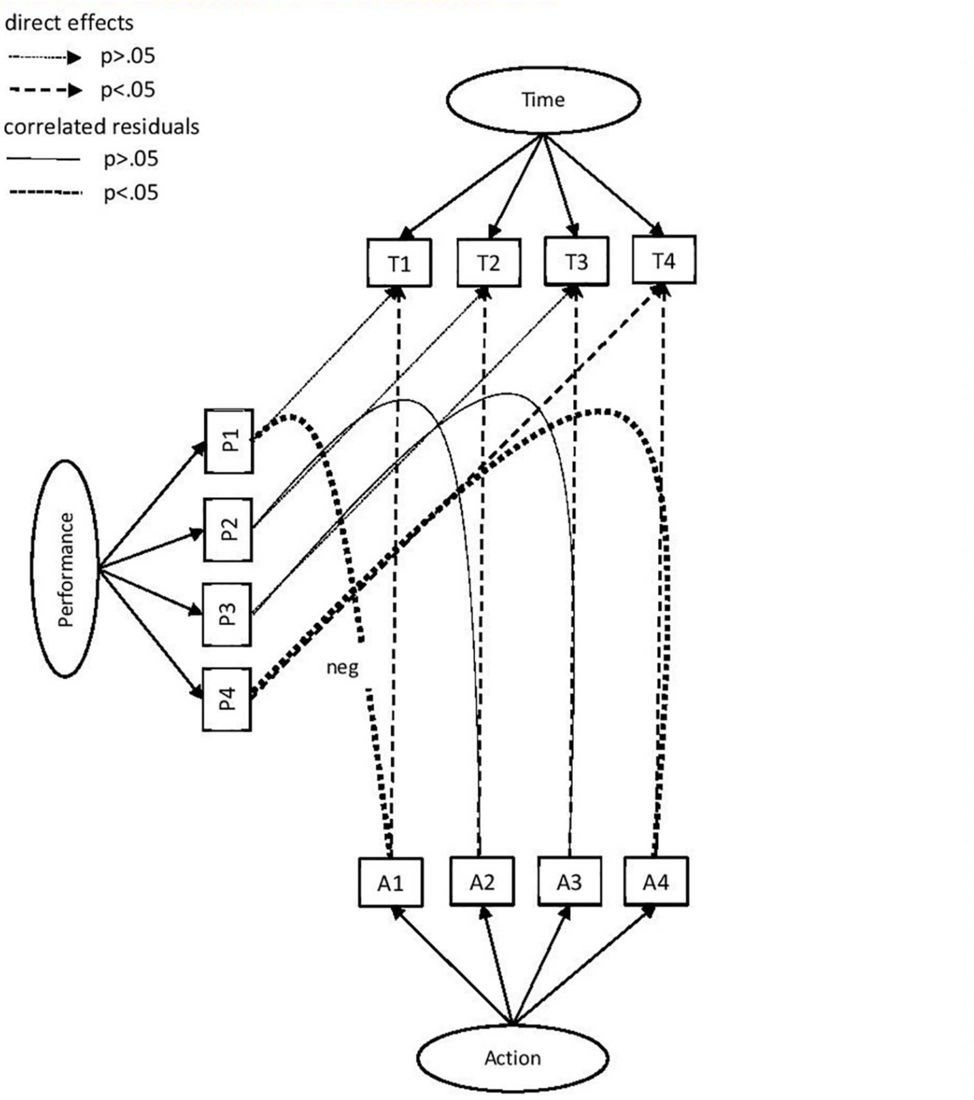


FIGURE 1 | Latent variable and dependency model for Xandar data. The latent variables are Time, Actions, and Performance. The observed variables per factor are indicated with capital letters referring to the latent variable (T, A, P) and with a number referring to the Xandar part (1, 2, 3, and 4). The direct effects between observed variables from the same Xandar part are indicated with single headed dashed arrows (between the A and T and between the P and T). The correlated residuals are indicated with dotted lines without arrow. Significance (p<–01) is denoted with a thicker dashed arrow (direct effects) or line (correlated residuals). All dependencies are positive except when indicated with "neg" (between Al and Pl). Correlations between latent variables, factor loadings, residual variances, and dependency values are omitted to avoid clutter in the figure. The correlations between the latent variables can be found in the text, the factor loadings are presented in **Table 3**, and the dependency values in **Table 4**.

DISCUSSION

For the model with loadings and dependencies showing in **Tables 3**, **4**, the latent variable correlations of Actions with Time and with Performance are negative. Hence, participants showing more activities are faster and perform less well in their collaborative problem solving. This is based on the United States dataset with the Xandar task. Successful participants take more time, perhaps a consequence of the previous two relationships. Multiplying the two negative correlations yields $-0.473 \times -0.723 = 0.346$, which is higher than the 0.190 estimate of the correlation between Time and Performance. This explains that in an alternative but formally equivalent model with an effect of Actions on Time and on Performance, the correlation between the residuals of the latent variables Time and Performance is negative. However, the correlation of -0.260 in question is not significant ($p > 0.05$).

The negative correlation between Actions and Time suggests that highly active students are fast and not so active students are slow. The combination of fast and active on the latent variables seem to reflect an impulsive and fast trial-and-error style. This strategy shows itself in the Xandar task as not very successful versus a slower, more thoughtful and apparently more successful style. It makes sense that respondents who are more deliberative may have more knowledge to bring to considering a successful solution, or be exhibiting more test effort in the Xandar context. We do not have the information to examine what is happening during the deliberation. This is in part because descriptions of the possible actions are not available in the data set. As well there is no interpretive information provided by PISA for the sample. This could include think-alouds where students describe why they are doing what they are doing. It could also have included qualitative response process information in which student explain their processes, in-depth interviews, or other approaches that supply interpretive information.

However, it makes of course sense that more actions take more time, which shows in the analysis of the dependencies between observed actions and time. This illustrates why it is informative to differentiate relationships between latent variables from relationships which show in dependencies.

Other important dependencies concern Part 4, which is a clearly reflective task, a kind of reflective and evaluative pause. The nature of the task may explain why performance is associated with more actions and requires more time, in contrast with Part 1 (agreeing on a strategy) where the association between actions and performance is negative. For instance, too much discussion on a strategy may signal a lack of structure.

For the result that the items examined can be considered as uni-dimensional although they are designed to tap on a diversity of processes and competencies, this suggests that the collaborative ability generalizes across processes. In other words, the collaborative competencies rely on a general underlying ability. The specificities of the processes are reflected in the extra dependencies. Part 4 involves monitoring and reflecting. This may explain why more activities and more time are associated with better performance. Part 1 by contrast involves planning and execution and representing and formulating. This may lead to better results if not based on trial and error (many actions) but on a structured and goal-oriented approach (less actions).

These dependencies suggest that, depending on the task, the collaborative ability may rely on a general underlying ability but be implemented through a different approach in various collaborative actions, as has been discussed in the literature (Fiore et al., 2017; OECD, 2017b; Eichmann et al., 2019). The special and specific status of Part 4 is also reflected in its lower loadings on all three latent variables (see standardized loadings).

Note that the extra dependencies here are not only considered for methodological reasons when variables stem from the same part. They may also reveal how subjects work on the tasks. This is consistent with the findings here. Parts such as 1 and 4 have a distinct theoretical description in the PISA framework. But how they draw on the collaborative ability can be seen in the empirical data to seemingly require different approaches as indicated in the process data.

Taken together, these results for the United States data set are consistent with problem solving performance modeled as invested time and number of actions.

Potential impacts underscore that it seems possible both to collect and to scale information on the collaborative ability. Measures may help provide intervention support, since in today's world especially, teams with good collaborative skills are necessary in any group. Groups can range from families to corporations, public institutions, organizations, and government agencies (OECD, 2013). Previously, dispositions to collaborate were reported based on the PISA data (Scalise et al., 2016). Indicators of collaborative ability also may be needed to create adequate interventions to train collaboration skills and to change current levels of individual collaboration.

As previously reported, the disposition dimensions of *collaborate*, *negotiate*, and *advocate/guide* might be useful starting points for creating such interventions (Scalise et al., 2016; OECD, 2017a). Alternatively, the factor structure here may yield suggestions on additional interesting starting points. This could include structures by which a student may approach collaboration (OECD, 2017b; Wilson et al., 2017) but more interpretive information would be needed. This could be combined with how participatory a student is disposed to be in collaboration, along with his or her team leadership inclinations, and beliefs in the value or efficacy of collaboration (Scalise et al., 2016).

Limitations to the analysis here include that only the United States data set of many countries available in the PISA data was analyzed. So this analysis should be extended to more countries and results compared in future work.

Also, from a statistical standpoint as discussed earlier, missing data were excluded listwise. In addition, minor but not major skewness was seen in two of the observed variables. Finally, multilevel modeling was not employed so the nested nature of students within schools was not taken into account.

TLI and RMSEA were reported here as the two fit indices since they seem most commonly used in the educational assessment field for large scale analyses. But there have been limited considerations for CPS on this topic.

For limitations from a conceptual standpoint, OECD releases a limited range of information, for instance items for only one of the 2015 collaborative problem solving tasks (Xandar) was released and collaborative actions were numbered but not described in the data set and data dictionary.

For implications of future work from this study, there are several. First, the era of analyzing process data and not only item response data in robust assessment tasks is upon us (many researchers including Praveen and Chandra, 2017). Approaches such as used here could be applied for other constructs, not just problem solving. Models can consider how to explore two types of relationship:

- at the level of general individual differences (the factors)
- at extra dependencies, which are direct effects and correlated residuals (independent of the factors)

These extra dependencies may provide a window on the underlying process dynamics, see **Figure 1**. It should be noted for implications for future work that it would be helpful if a range of simplified visualizations could be developed for such complex analyses. Standard plots after including dependencies seemed too complex to be fully useful.

For extensions to the specific modeling here, it would be important as discussed earlier to explore fitting the same or similar models across data sets from other countries (Thomas and Inkson, 2017). This could be augmented by also modeling potential country-level effects at the item level, by exploring differential item functioning. Furthermore it would be interesting to consider covariates available in the PISA student questionnaire data set (SQ) in relation to the collaborative ability examined here. This could include indicators for dispositions for collaborative problem solving that moved forward to the main PISA study (Scalise et al., 2016). These indicators include student-level indicators available in the CPS SQ data set regarding self-report of dispositions toward cooperation, guiding, and negotiating.

It should also be mentioned that other very interesting student-level indicators regarding additional preferences in collaboration had to be dropped from the PISA main study. This was due to time limitations. Dropped indicators included dispositions toward collaborative *leadership*, as well as student-level indicators of in-school and out-of-school collaborative *opportunities*. While these were not possible to include in the main study due to time limitations for the PISA administration, the indicators were part of the field testing. They could be very interesting to administer at the country-level in other national or international assessments.

Teacher-level indicators are also available in the PISA data set that provide information on opportunity to learn (OtL) for students in the PISA CPS. Data include classroom-level OtL reports of team activities, grouping practices, types of collaborative activities, and types of rewards provided for engaging in successful team work. Exploring relationships here might allow more reflection on connections to potential interventions. The PISA data are cross-sectional but might help to inform research studies within countries.

In closing, it is important to mention that the creation and delivery of the innovative PISA CPS instrument included both simulated collaboration of a hard-to-measure construct (Scalise, 2012) and sharing of some process data. This was critical to the examination here, as has been the case for other collaboration-oriented assessments (Greiff et al., 2014, 2015, 2016). This analysis underscores that addressing challenges of education in the 21st century may continue to require new data sources, to address new challenges for education worldwide.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01280/full#supplementary-material

REFERENCES

Adejumo, G., Duimering, R. P., and Zhong, Z. (2008). A balance theory approach to group problem solving. *Soc. Netw.* 30, 83–99. doi: 10.1016/j.socnet.2007.09.001

Avouris, N., Dimitracopoulou, A., and Komis, V. (2003). On evaluation of collaborative problem solving: methodological issues of interaction analysis. *J. Comput. Hum. Behav.* 19, 147–167.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., et al. (2012). "Defining twenty-first century skills," in *Assessment and Teaching of 21st Century Skills*, eds P. Griffin, B. McGaw, and E. Care (New York, NY: Springer).

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., and Rumble, M. (2010). "Assessment and teaching of 21st century skills: defining 21st century skills," in *White Paper released at the Learning and Technology World Forum* 2010, (London).

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2002). *Applied Multiple Regression/Correlationanalysis for the Behavioral Sciences*, 3rd Edn. Hove: Psychology Press.

Cooke, N. J., Kiekel, P. A., Salas, E., Stout, R., Bowers, C., and Cannon- Bowers, J. (2003). Measuring team knowledge: a window to the cognitive underpinnings of team performance. *Group Dyn.* 7, 179–219.

Eichmann, B., Goldhammer, F., Greiff, S., Pucite, L., and Naumann, J. (2019). The role of planning in complex problem solving. *Comput. Educ.* 128, 1–12. doi: 10.1016/j.compedu.2018.08.004

Fiore, S. M., Graesser, A., Greiff, S., Griffin, P., Gong, B., Kyllonen, P., et al. (2017). *Collaborative Problem Solving: Considerations for the National Assessment of Educational Progress*. Available at: https://nces.ed.gov/nationsreportcard/pdf/researchcenter/collaborative_problem_solving.pdf (accessed April 11, 2018).

Foltz, P. W., and Martin, M. J. (2008). "Automated communication analysis of teams," in *Team Effectiveness in Complex organisations and Systems: Cross-Disciplinary Perspectives and Approaches*, eds E. Salas, G. F. Goodwin, and S. Burke (Boca Raton, FL: CRC Press).

Graesser, A. C., Jeon, M., and Dufty, D. (2008). Agent technologies designed to facilitate interactive knowledge construction. *Discourse Process.* 45, 298–322. doi: 10.1080/01638530802145395

Greiff, S., Krkovic, K., and Hautamäki, J. (2016). The prediction of problem solving assessed via microworlds: the relative importance of fluid reasoning and working memory. *Eur. J. Psychol. Assess.* 32, 298–306. doi: 10.1027/1015-5759/a000263

Greiff, S., Wüstenberg, S., and Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Comput. Educ.* 91, 92–105. doi: 10.1016/j.compedu.2015.10.018

Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, J., Graesser, A., et al. (2014). Domain-general problem solving skills and education in the 21st century. *Educ. Res. Rev.* 13, 74–83. doi: 10.1016/j.edurev.2014.10.002

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., and Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: a clarification of the literature. *J. Appl. Psychol.* 86, 730–740. doi: 10.1037/0021-9010.86.4.730

OECD (2013). *PISA 2015: Draft Collaborative Problem Solving Framework*. Available at: http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf (accessed September 26, 2014).

OECD (2014). *PISA 2012 Results: Creative Problem Solving: Students' Skills in Tackling Real-Life Problems (Volume V)*. Paris: OECD.

OECD (2016). *Description of the Released Unit from the 2015 PISA Collaborative Problem-Solving Assessment, Collaborative Problem-Solving Skills, and Proficiency Levels*. Available at: https://www.oecd.org/pisa/test/CPS-Xandar-scoring-guide.pdf (accessed December 10, 2018).

OECD (2017a). *Chapter 17: Questionnaire Design and Computer-based Questionnaire Platform. In PISA 2015 Technical Report*. Available at:

http://www.oecd.org/pisa/data/2015-technical-report/ (accessed September 21, 2017).

OECD (2017b). *PISA 2015 Results (Volume V): Collaborative Problem Solving, PISA*. Paris: OECD Publishing.

O'Neil, H. F., Chung, G., and Brown, R. (1997). "Use of networked simulations as a context to measure team competencies," in *Workforce readiness: Competencies and assessment*, ed. H. F. O'Neil (Mahwah, NJ: Lawrence Erlbaum Associates), 411–452.

Praveen, S., and Chandra, U. (2017). Influence of structured, semi-structured, unstructured data on various data models. *Int. J. Sci. Eng. Res.* 8, 67–69.

R Core Team (2018). *R: A language and environment for statistical computing*. Vienna: R Core Team.

Rosseel, Y. (2012). lavaan: an r package for structural equation modeling. *J. Stat. Softw.* 48, 1–36. doi: 10.3389/fpsyg.2014.01521

Rummel, N., and Spada, H. (2005). Learning to collaborate: an instructional approach to promoting collaborative problem solving in computer-mediated settings. *J. Learn. Sci.* 14, 201–241. doi: 10.1207/s15327809jls1402_2

Scalise, K. (2012). "Using technology to assess hard-to-measure constructs in the CCSS and to expand accessibility," in *Proceedings of the Invitational Research Symposium on Technology Enhanced Assessments*, (Princeton, NJ).

Scalise, K., and Binkley, M. (2009). "Transforming educational practice by transforming assessment: update on assessment & teaching of 21st century skills," in *PISA Problem Solving 2012*, (Santa Barbara, CA).

Scalise, K., Mustafić, M., and Greiff, S. (2016). "Dispositions for collaborative problem solving," in *Assessing Context of Learning World-Wide (Methodology of Educational Measurement and Assessment Series)*, eds S. Kuger, E. Klieme, N. Jude, and D. Kaplan (Dordrecht: Springer).

Thomas, D. C., and Inkson, K. (2017). "Communicating and negotiating across cultures," in *Cultural intelligence: Surviving and Thriving in the Global Vilage*, 3rd Edn, (Oakland, CA: Berrett-Koehler), 76–97.

Vogel, F., Wecker, C., Kollar, I., and Fischer, F. (2016). Socio-cognitive scaffolding with computer-supported collaboration scripts: a meta-analysis. *Educ. Psychol. Rev.* 29, 477–511. doi: 10.1007/s10648-016-9361-7

Wilson, M., Gochyyev, P., and Scalise, K. (2017). Modeling data from collaborative assessments: learning in digital interactive social networks. *J. Educ. Meas.* 54, 85–102. doi: 10.1111/jedm.12134

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Psychology

ORIGINAL RESEARCH
published: 10 June 2019
doi: 10.3389/fpsyg.2019.01368

Making the Psychological Dimension of Learning Visible: Using Technology-Based Assessment to Monitor Students' Cognitive Development

Gyöngyvér Molnár[1] and Benő Csapó[2]*

[1] Department of Learning and Instruction, University of Szeged, Szeged, Hungary, [2] MTA–SZTE Research Group on the Development of Competencies, University of Szeged, Szeged, Hungary

Edited by:
Ronny Scherer,
University of Oslo, Norway

Reviewed by:
Alison Margaret Gilmore,
University of Otago, New Zealand
Andreas Rausch,
Universität Mannheim, Germany

***Correspondence:**
Gyöngyvér Molnár
gymolnar@edpsy.u-szeged.hu

Specialty section:
This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Psychology

Received: 07 December 2018
Accepted: 27 May 2019
Published: 10 June 2019

Citation:
Molnár G and Csapó B (2019)
Making the Psychological Dimension
of Learning Visible: Using
Technology-Based Assessment
to Monitor Students' Cognitive
Development.
Front. Psychol. 10:1368.
doi: 10.3389/fpsyg.2019.01368

Technology-based assessment offers unique opportunities to collect data on students' cognitive development and to use that data to provide both students and teachers with feedback to improve learning. The aim of this study was to show how the psychological dimension of learning can be assessed in everyday educational practice through technology-based assessment in reading, mathematics and science. We analyzed three related aspects of the assessments: cognitive development, gender differences and vertical scaling. The sample for the study was drawn from primary school students in Grades 1–8 (ages 7 to 14) in Hungary. There were 1500 to 2000 students in each grade cohort. Online tests were constructed from 1638 items from the reading, mathematics, and science domains in the eDia system. The results confirmed that the disciplinary, application and psychological dimensions of learning can be distinguished empirically. Students' cognitive development was the most steady (and effective) in mathematics, where the greatest development occurred in the first years of schooling. Path models suggested that the psychological dimension of learning can be predicted at a moderate level based on students' level of school knowledge consisting of the disciplinary and application dimensions of learning as latent constructs. The predictive power was almost the same in both dimensions. Generally, girls developed faster in the psychological dimension of reading, mathematics and science learning; however, the size of gender differences varied by age and domain. This study (1) provides evidence that the psychological dimension of learning can be made visible even in an educational context, (2) highlights the importance of the explicit development of the psychological dimension of learning during school lessons, and (3) shows that there are gender differences in the developmental level of the psychological dimension of learning in favor of girls but that this varies by grade and domain.

Keywords: technology-based assessment, online assessment, assessment for learning, visible learning, cognitive development

INTRODUCTION

Improving students' cognitive abilities has always been a goal of schooling since the very beginning of formalized education (Hattie and Anderman, 2013). However, despite the theoretical foundations, assessment instruments and pedagogical practices that have evolved over time, this aim has not yet been met; in many school systems students' cognitive abilities are not optimally enhanced. In the 20th century, several research schools and paradigms sought to conceptualize cognition, define its key constructs and make them measurable (see e.g., Binet and Simon, 1916; Inhelder and Piaget, 1958; Adey, 2007). Among these, research on intelligence and the related psychometric tradition, Piaget and his school, and the cognitive revolution have all had a major impact on redefining the goals of education. The implications of the research within these paradigms were drawn for educational practice, and a number of mostly stand-alone programs were initiated in the 1970s, outside classroom instruction (Feuerstein et al., 1980; Klauer, 1989a,b, 1991, 1993, 1997). Later on, in the 1990s, developmental effects were embedded in school subjects using the content of learning (Adey and Shayer, 1994; Shayer, 1999; Adey et al., 2001; Shayer and Adey, 2002; Shayer and Adhami, 2007). The related research, including a number of experiments, resulted in a better understanding of the role that cognitive processes play in school learning, but it has had a modest impact on educational practice.

At the beginning of this millennium, more or less the same ideas emerged in a new wave of teaching 21st-century skills. Several projects sought to define, operationalize, measure and teach these skills (see e.g., Trilling and Fadel, 2009; Griffin and Care, 2014), but the same constraints appeared to hinder progress in putting these ideas into practice in mass education as with previous similar attempts. There were no proper tools for assessing and monitoring changes in students' cognition. The availability of appropriate assessment instruments is a necessary condition for any pre-test – post-test experimental design as well. However, what can be created and applied in specific experimental conditions cannot always be scaled up for broader practical applications. Similarly, the roots of a number of practical educational challenges can be traced back to the fact that significant determinants of school learning are not visible (Hattie, 2009). They are also not easy to observe, nor can developmental deficiencies always be identified by teachers (MacGilchrist et al., 2004). The lack of thinking skills – the cognitive tools required for successful learning – are not identified; thus, they remain untreated, and this significantly hampers further learning.

Thinking, or more specifically, a set of cognitive skills essential for learning, such skills are not observable in the everyday educational context. Students are not aware of the existence of the required processes, and teachers, even if they receive training in identifying the cognitive processes underlying learning, are not able to observe them, or they simply have no time or capacity to determine each student's individual needs. Although the developmental levels of crucial thinking skills might be measured with traditional paper-based instruments, the immense costs, the human resources required, and the time between assessment and feedback excludes the possibilities of using them diagnostically. Technology may be a solution for making thinking processes visible by creating simpler, faster, frequently applicable and cost-effective assessments (Mayrath et al., 2012).

In this paper, which is part of a larger project, we present the results of work in identifying cognitive processes relevant for learning, making them measurable in normal educational contexts, and providing students and teachers with frequent feedback. One of the most challenging aspects of this work, is establishing the validity of diagnostic instruments to assess of cognitive processes; showing that the tests measure something more than mastering the current teaching material. To do this, we empirically validated a 3-dimensional framework developed for diagnostic assessment and explored the psychometric characteristics of an item bank devised for the assessment of the psychological dimension of learning.

THEORETICAL BACKGROUND

The idea of making learning visible was introduced into educational research and development by John Hattie. He made a great step forward in initiating evidence-based educational practice when he synthesized the results of over 800 meta-analyses (Hattie, 2009). He translated his findings into actual classroom work, and in his book for teachers, he explained:

> Visible teaching and learning occurs when learning is the explicit and transparent goal, when it is appropriately challenging, and when the teacher and the student both (in their various ways) seek to ascertain whether and to what degree the challenging goal is attained, when there is deliberate practice aimed at attaining mastery of the goal, when there is feedback given and sought, and when there are active, passionate, and engaging people (teachers, students, peers, and so on) participating in the act of learning (Hattie, 2012, p. 18).

As he emphasizes, feedback plays a central role in successful learning, which at a higher level of learning, includes self-monitoring, self-evaluation and self-assessment. However, he also explains how difficult a task it is to provide proper feedback: "Learners can be so different, making it difficult for a teacher to achieve such teaching acts: students can be in different learning places at various times, using a multiplicity of unique learning strategies, meeting different and appropriately challenging goals" (Hattie, 2012, p. 18).

Student diversity, i.e., students at different levels in different cognitive attributes, is not the most challenging phenomenon when proper feedback is considered. A major problem is that a number of learning outcomes, sometimes the most important ones, are not visible and cannot easily be made visible. While the majority of the studies Hattie reports on deal with organizational issues, methods and classroom practices for teaching curricula, there are far fewer studies that cover the underlying cognitive processes, e.g., reasoning skills, required to understand mathematics and science or precursors of reading, such as phonemic awareness. Some studies have focussed on the most hidden aspects of learning. For example, Ritchhart et al. (2011) identify a broad range of teaching and learning practices to make thinking visible. They identify the crucial problem in a simplified conception of learning (reduced to memorization) and knowledge (reduced to information, facts and figures): "When we demystify the thinking and learning processes, we provide models for students of what it means to engage with ideas, to think, and to learn. In doing so, we dispel the myth that learning is just a matter of committing the information in the textbook to one's memory" (Ritchhart et al., 2011, p. 28).

Taking into account diversity among students, the limited capacity of teachers and the need to provide feedback on the most relevant but least visible aspects of school learning – promoting students' cognitive development – we may conclude that students and teachers need a different approach to assessment to improve learning. The online assessment system, eDia, was designed for this purpose. It assesses "thinking," or "cognitive development," as a separate dimension, which we call the psychological dimension of learning. We briefly introduce the 3-dimensional theoretical framework that forms the basis for the diagnostic assessment system, and then we elaborate on the psychological dimension in more detail, as that is the focal topic of the present study. Finally, we discuss the crucial role of technology, arguing that its widespread availability in schools makes the time right for such a system to be introduced and integrated into regular educational processes.

Learning and Cognitive Development: A 3-Dimensional Model of Learning Outcomes

An online diagnostic assessment system, eDia, has been constructed to provide teachers and students with relevant feedback information (Csapó and Molnár, unpublished). The eDia system covers the three most frequently assessed domains of school education; reading, mathematics and science. Large item banks have been developed for use in regular classroom assessments in Grades 1 to 6 of primary school, and for Grades 7 and 8 to explore the developmental trends in a broader age range.

The objectives of each item bank are defined in its assessment framework, similarly to international comparative studies, such as Trends in International Mathematics and Science Study (TIMSS; Mullis et al., 2005) and Progress in International Reading Literacy Study (PIRLS; Mullis and Martin, 2015); they are based on a 3-dimensional model of the goals of learning that forms a common foundation for diagnostic assessment. The three dimensions include thinking/reasoning, application and disciplinary knowledge. [The 3-dimensional framework has been published in several articles and book chapters before the assessment frameworks were elaborated (see e.g., Csapó, 2010; Nunes and Csapó, 2011; Adey and Csapó, 2012; Blomert and Csépe, 2012)]. The framework for reading was somewhat different those for mathematics and science (Csapó and Csépe, 2012; Csapó et al., 2015c), which were more similar (Csapó and Szendrei, 2011; Csapó and Szabó, 2012; Csapó et al., 2015a,b).

The intention of "cultivating the mind" – developing cognitive abilities – may be traced back to ancient philosophy. To set goals in this direction, a model of mind is needed; more specifically, knowledge of how internal psychological attributes are structured and how psychological processes play a role in learning (see more details in the next section). In the eDia frameworks, this is the "thinking" (this term is mostly used in the context of mathematics and science), or, more generally, the "psychological dimension." According to the model, we propose the psychological dimension of knowledge does not only contain "domain-specific reasoning skills," but also general reasoning skills embedded in different content and contexts, which has lately been referred to as transversal skills; and is not the same as procedural knowledge. We assume that there are natural cognitive developmental (psychological) processes. These processes, as described by Piaget, take place in the interaction between the child and his/her environment. School education may stimulate this development if it provides a student with proper environmental stimuli and if these stimuli are within the zone of proximal development (ZPD) of the child (Vygotsky, 1978). Very often, school instruction is not adjusted to the individual needs of the students; usually the stimuli are far beyond their ZPD. In these cases, students benefit little from instruction; they memorize the rules and develop specific skills through a large amount of drill practice, which have any real impact on their cognitive development. For example, students may learn rules to deal with ratios and proportions without this learning having much impact on the development of proportional reasoning. Schools may teach students a great deal about combinatorics, probability and correlation without having a real impact on the development of combinatorial, probabilistic or correlative reasoning. In this way, we distinguish the psychological dimension from the disciplinary dimension, which may include procedural knowledge (e.g., skills for solving linear equations or proving geometric theorems) or domain-specific reasoning skills. This model and approach opens the door to fostering domain-general reasoning skills in a domain-specific context.

Application deals with another ancient goal – that school should teach something that is applicable beyond the school context. Applying knowledge and transferring it to new contexts require a deeper conceptual understanding and usually specific exercises to facilitate application. Therefore, most knowledge mastered at school remains inert and not applicable in new contexts (Alexander and Murphy, 1999; Bransford and Schwartz, 1999; Csapó, 2010). The PISA conducted by the OECD has focussed on this dimension from the very beginning. The PISA expert groups elaborated the concept of applicable knowledge and defined it as competencies students need in a modern society. To develop such a framework, the social relevance of knowledge, i.e., the needs of societies have also be taken into account. For the frameworks of the first and second PISA assessment, the concept of literacy was extended in include the objects of the assessment in the three domains as reading literacy, mathematical literacy and scientific literacy (OECD, 1999, 2003).

Disciplinary knowledge is the third dimension and is most commonly known as curricular content. Arts and sciences content constitutes the major source of disciplinary knowledge. The first major international comparative studies (e.g., First and Second International Mathematics Study – Husén, 1967;

Burstein, 1993; First and Second International Science Study – Bloom, 1969; IEA, 1988), the precursor to the TIMSS, assessed this dimension. The first assessments were based on an analysis of the curricula in the participating countries. More recently, the TIMSS frameworks organize the objects of the assessment into three groups: content, application and reasoning. This classification bears some similarity to the 3-dimensional eDia frameworks (For PISA assessment frameworks, see OECD, 2003).

Education must not be reduced to providing the right answer quickly, but must deal with the ongoing cognitive work of understanding new ideas and information that will serve students as learners in the future (Costa and Kallick, 2009). In modern society, students are expected to apply their knowledge in a wide range of contexts, and they should be able to solve problems in unknown, novel situations. Thus, these goals must reinforce and interact with each other as they are strongly connected (Molnár and Csapó, 2019).

It is reasonable that the earliest efforts to measure knowledge learnt at school focussed on areas that were the easiest to measure: the disciplinary (knowledge) dimension of learning (see e.g., IEA TIMSS). The goal of applying that knowledge in a new context (the application dimension) and assessing students' ability to do so is a more complicated task (see e.g., OECD PISA). The goal of developing students' thinking abilities (the psychological dimension) is even more complex. To be able to make thinking visible, we must be clear about, and draw on, our understanding of what thinking is and what types of thinking we want to assess and enhance.

Assessment Beyond the Content of Actual Learning

In the 20th century, several research paradigms have conceptualized the development of thinking and its relationship to school education. Among these, research on "intelligence" was the first that was closely linked to education. The first intelligence test (Binet–Simon test, Binet and Simon, 1916) was constructed to assess children's preparedness for schooling, and the Scholastic Aptitude Test (SAT) (see Grissmer, 2000) served a similar purpose at the transition from secondary to tertiary education. Several new approaches, models and interpretations of the concept of intelligence have been proposed. From the perspective of education, the more useful ones consider intelligence as (able to be modified, taught, learnt, or improved within educational contexts). Our psychological dimension in each domain may thus overlap with the inductive reasoning components of "fluid" intelligence. The psychological dimension can be embedded within the conception of plastic general ability (see Adey et al., 2007), and a number of cognitive skills covered by the psychological dimension of our frameworks are explicitly identified in Carroll's three-strata model of abilities (Carroll, 1993) and the Specialized Cognitive Systems of Demetriou's model (Demetriou et al., 1992, 1993; Adey et al., 2007). On the other hand, we emphasize that all the cognitive skills discussed in the psychological dimension of the frameworks are embedded within the content and context of each particular domain, and the tasks developed from the frameworks are adjusted to the developmental level of the cohort of students to be assessed.

The work of Jean Piaget and his school was characterized by another approach. Piaget described students' reasoning skills with well-defined operations, which correspond with certain mathematical structures (see e.g., Inhelder and Piaget, 1958). He mostly used basic science content for his experiments (e.g., the pendulum), and the operations he identified may be found in various learning contexts as well as in everyday problems.

The cognitive revolution in psychology provided a new impetus to research efforts in school learning. It led to a more differentiated conception of knowledge and learning, allowing a more precise definition of the goals of education. Recent studies in psychology and education have shown that these skills are especially crucial at the beginning years of schooling, as students' developmental level determines later success (see Nguyen et al., 2016).

The psychological dimension has been conceptualized as the interaction between the development of students' thinking skills and learning at school (Nunes and Csapó, 2011; Adey and Csapó, 2012; Blomert and Csépe, 2012) and must address how students learn in reading, mathematics and science.

In this study, we explored the prospects of making the psychological dimension of learning visible by using technology-based assessments to monitor the development of students' thinking skills. The aim of this study was to show how the psychological dimension of learning (thinking) can contribute to the development of specific reasoning skills.

In reading, assessment of the psychological dimension (thinking and reasoning) covers the cognitive mechanisms of development from laborious phonological decoding to the automatic recognition of whole words, and from prerequisite skills of reading through phonemic, phonological and morphological awareness to metacognitive aspects (Blomert and Csépe, 2012). In mathematics (Nunes and Csapó, 2011) and science (Adey and Csapó, 2012), there are generic objects and domain specific objects. For example, number sense is specific to mathematics, while the control of variables and scientific reasoning are better covered within the science framework. Operational reasoning (e.g., seriation, class inclusion, classification, combinatorial reasoning, probabilistic reasoning, proportional reasoning) and some higher-order thinking skills (e.g., inductive reasoning and problem solving) are more generic and can be assessed in both mathematics and science.

AIMS, RESEARCH QUESTIONS AND HYPOTHESES

In this study, we explored the prospects of making the psychological dimension of learning visible by using technology-based assessments to monitor the development of students' reasoning skills. The aim of the study was to show how the psychological dimension of learning (thinking) can be assessed in everyday educational practice and how it is related to students' level of subject matter content knowledge. Three domains were explored from this perspective: reading, mathematics and science. Reading is the basis for all further learning, including mathematics and science, while mathematics provides foundations for learning in various areas of science. These domains are central in many education systems, and large-scale international comparative studies, such as TIMSS, PIRLS, and PISA, have focussed on these areas. We analyzed three aspects of the assessments: cognitive development, gender differences and vertical scaling.

Worldwide, there are many initiatives and computer-based tests available in the domains of reading, mathematics and science worldwide. However, they mainly focus only on disciplinary knowledge dimension (content) or the application dimension (literacy of learning) (e.g., TIMSS, PIRLS, and OECD PISA). There are no regular large-scale assessments that include the psychological dimension of learning in primary school – cognitive development. The available assessment systems in reading, mathematics and science have been designed to assess older students' reading, mathematics and science knowledge (e.g., TIMSS, PIRLS, and PISA). The present study sought to: (1) define and examine the different dimensions of learning in reading, mathematics and science; (2) monitor and compare cognitive development (the psychological dimension of learning) in the three domains over time; (3) analyze the proportion of unexplained variance in cognitive development if school knowledge (the application and disciplinary dimensions) is taken into account in reading, mathematics and science; and (4) identify any gender differences in the cognitive development in the three domains. We sought to answer five research questions.

RQ1: Can the three dimensions of learning be distinguished empirically? We explored this question to see if cognitive development, the development of reasoning skills, can be assessed separately and be made visible in everyday educational practice. We hypothesize that the psychological, application and disciplinary dimensions of learning can be distinguished empirically, assessed and monitored in everyday educational practice (Csapó and Szendrei, 2011; Csapó and Csépe, 2012; Csapó and Szabó, 2012). We also hypothesize that they will interact and correlate with each other.

RQ2: Is the psychological dimension of learning the same across the three domains? That is, is the same construct being measured in the psychological dimension of learning across the three main domains? The roots of cognitive development may be universal as early neurocognitive development in children is similar across cultures and societies (Molnár and Csapó, 2019). Therefore, based on the conceptualization of the psychological dimension of learning as the interaction between students' cognitive development and learning at school (Nunes and Csapó, 2011), we hypothesize that the 1-dimensional model will fit the data better than the 3-dimensional model. However, we argue that the 3-dimensional model will take into account results from research on knowledge transfer. According to McKeachie (1987), "Spontaneous transfer is not nearly as frequent as one would expect" (p. 709).

RQ3: How does the psychological dimension of reading, mathematics and science develop over time during primary schooling? Based on previous research results on reasoning skills, we hypothesize that children's cognitive development is slow (Molnár et al., 2013, 2017), indicating the need for more stimulating school lessons. Based on Polya's (1981) theory of problem solving, and results from research on mathematics teaching (e.g., Nunes and Csapó, 2011), we hypothesize that the psychological dimension of learning in mathematics will develop the most readily.

RQ4: How can the psychological dimension of learning be explained by students' level of school knowledge in reading, mathematics and science? That is, how can learning in reading, mathematics and science contribute to the development of the psychological dimension of learning, and how effectively does it stimulate students' general cognitive development? Research in this field provides rich resources ranging from the classical work of Piaget (see e.g., Inhelder and Piaget, 1958) to the most recent neurocognitive studies (such as Geake and Cooper, 2003; Thomas et al., 2019). We hypothesize that learning reading, mathematics and science will contribute to students' development in the psychological dimension of learning but that the transfer effect will be low. We base our hypothesis on empirical research that has found that reasoning skills develop relatively slowly during primary and secondary education with the average pace of development being about one quarter of a standard deviation per year (Csapó, 1997; Molnár and Csapó, 2011; Greiff et al., 2013; Molnár et al., 2013, 2017). The development of reasoning skills is a "by-product" of teaching rather than guided by explicit instruction (de Koning, 2000).

RQ5: How does the developmental level of the psychological dimension of learning differ by gender, grade and domain? Based on the most prominent international studies (Martin et al., 2016; Mullis et al., 2016, 2017; OECD, 2016) and the research results on gender differences in students' development of reasoning skills (Wüstenberg et al., 2014), we hypothesize gender differences in the development of the psychological dimension of learning will vary by grade and domain. The PISA studies indicated that the achievement of 15-year-old Hungarian girls in the application dimension of reading was significantly better than that of boys, while there were no statistically significant gender differences in mathematics and science (OECD, 2016). In contrast, the TIMSS studies that focus on younger students (Grades 4 and 8; 10- to 14-year-olds) mainly assess the disciplinary dimension of mathematics and science knowledge. Their findings indicated that boys significantly outperform girls in mathematics in Grade 8 (Mullis et al., 2016), but there was no statistically significant gender difference in Grade 4. In science, boys significantly outperformed girls at both grade levels (Martin et al., 2016). In PIRLS, Grade 4 Hungarian girls significantly outperformed their boys in reading (Mullis et al., 2017). Please note that the present study focussed on the psychological dimension and not on the application or disciplinary dimensions of learning in mathematics, science or reading.

MATERIALS AND METHODS

Participants

The sample of students for the study was chosen from the partner school network of the Center for Research on Learning and Instruction at the University of Szeged in Hungary. As schools participated voluntarily in the project, representative sampling of school classes or students was not a goal. However, based on the data collected from the schools, it was possible to generate nationally representative indicators for the main variables. We noted that schools with relatively large numbers of low socioeconomic (SES) students were under-represented in the present study, possibly due to the lack of ICT available in those schools.

The sampling unit was a school class. Classes were drawn from primary and secondary schools from Grades 1–8 (aged 7–14). A total of 656 classes from 134 schools in different regions were involved in the study, resulting in a wide-ranging distribution of students' background variables. The total number of students involved in the study was 14,062 (**Table 1**). The proportion of boys and girls was about the same. As participation was voluntary, not all students completed tests in all three domains or in each dimension within each domain. Thus, data was potentially available for students who completed nine elements: the assessment of three dimensions of learning (psychological, application, and disciplinary) in three domains (reading, mathematics, and science). After the scaling procedure, we excluded students from the analyses where, because of missing data. it was not possible to compute an ability level in at least one of the nine elements. Thus, 5,714 students from 310 classes and 97 schools were involved in the analyses.

Tests

An item bank was constructed for diagnostic assessments in reading, mathematics and science based on the three dimensions of learning described in the previous section. These item banks collectively contained almost 17,000 tasks with most tasks having several items. There were 6685 tasks for reading, 6691 for mathematics and 3535 for science. Tests to measure the psychological, application, and disciplinary dimensions of learning in reading, mathematics and science among students in Grades 1–6 (aged 6–7 to 12–13). The tests for the study were drawn from these item banks. Students in Grades 7 and 8 received tasks originally written for students in Grades 5 and 6 (see **Table 2**).

For each grade level, nine tasks with different difficulty levels (three easy, three medium-difficulty and three difficult) were chosen from each item bank to assess each dimension. After this procedure, there were 543 tasks in reading, 604 in mathematics and 492 in science.

The tasks were grouped into clusters, with 10–15 items per cluster for students in the lower grades and 15–20 items for students in the higher grades. One 45-min test consisted of four clusters of tasks for students in Grades 1 and 2 (50–55 items) and five clusters for students in Grades 3 to 6 (60–85 items). Each test contained clusters of tasks from each learning dimension with the clusters positioned in a different order to avoid the item-position effect in the scaling procedure. Anchor items were used within and between the different grades for the horizontal and vertical scaling of the data. The clusters contained easier or harder tasks from lower or higher grades. A total of 483 strongly anchored, but different clusters were developed from the items selected.

For optimizing the measurement error of the test, the clusters contained tasks from the same dimension of learning, ranging in task difficulty for the different grade levels. That is, students received more tasks from one learning dimension if those tasks were originally prepared for students in lower or higher grades. The structure of the test of mathematical knowledge is presented in **Table 2** paralleled the structure of the reading and science tests. Based on this structure, 162 different tests (nine in each grade and each domain) were constructed from the item banks for the vertical scaling of students in Grades 1–8.

TABLE 1 | The sample for the study.

Grade	Whole sample			Data analyzed (3 domains × 3 dimensions)		
	N	Age [mean (SD)]	Gender (% of girls)	N	Age [mean (SD)]	Gender
1	1003	7.8 (0.58)	47.2	349	7.8 (0.59)	46.3
2	1348	8.8 (0.61)	51.5	528	8.8 (0.57)	49.6
3	1675	9.8 (0.62)	49.9	598	9.8 (0.65)	49.9
4	2148	10.8 (0.60)	50.2	659	10.8 (0.60)	49.3
5	2441	11.8 (0.60)	47.8	1169	11.8 (0.61)	47.8
6	2122	12.9 (0.59)	47.7	1017	12.9 (0.59)	47.0
7	1875	13.9 (0.62)	49.6	800	13.9 (0.61)	50.0
8	1450	14.9 (0.63)	49.5	594	14.9 (0.63)	49.7
Total	14062	11.6 (2.18)	49.1	5714	11.7 (2.15)	49.0

TABLE 2 | The structure of the tests in mathematics by cluster of tasks for each grade level.

Grade	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1	Mouse usage warm-up tasks	MD1 (15)	MR1 (15)	MA1 (15)	MD2/MR2/MA2 (10)
		MA1 (15)	MD1 (15)	MR1 (15)	MD2/MR2/MA2 (10)
		MR1 (15)	MA1 (15)	MD1 (15)	MD2/MR2/MA2 (10)
2	Mouse usage warm-up tasks	MD1 (15)	MR2 (15)	MA2 (15)	MD3 (10)
		MA1 (15)	MD2 (15)	MR2 (15)	MA3 (10)
		MR1 (15)	MA2 (15)	MD2 (15)	MR3 (10)
3	MD1 (10)	MD2 (10)	MA2 (15)	MR3 (15)	MD4 (10)
	MA1 (10)	MA2 (10)	MR2 (15)	MD3 (15)	MA4 (10)
	MR1 (10)	MR2 (10)	MD2 (15)	MA3 (15)	MR4 (10)
4	MD2 (10)	MD3 (10)	MA4 (15)	MR4 (15)	MD5 (10)
	MA2 (10)	MA3 (10)	MR4 (15)	MD4 (15)	MA5 (10)
	MR2 (10)	MR3 (10)	MD4 (15)	MA4 (15)	MR5 (10)
5	MD3 (15)	MD4 (15)	MA5 (20)	MR5 (20)	MD6 (15)
	MA3 (15)	MA4 (15)	MR5 (20)	MD5 (20)	MA6 (15)
	MR3 (15)	MR4 (15)	MD5 (20)	MA5 (20)	MR6 (15)
6	MD4 (15)	MD5 (15)	MA6 (20)	MR6 (20)	MD6 (15)
	MA4 (15)	MA5 (15)	MR6 (20)	MD6 (20)	MA6 (15)
	MR4 (15)	MR5 (15)	MD6 (20)	MA6 (20)	MR6 (15)
7–8	MD5 (15)	MR5 (15)	MA5 (20)	MD6 (20)	MR6 (15)
	MA5 (15)	MD5 (15)	MR5 (20)	MA6 (20)	MD6 (15)
	MR5 (15)	MA5 (15)	MD5 (20)	MR6 (20)	MA6 (15)

M, mathematics; D, disciplinary dimension; A, application dimension; R, reasoning dimension; 1–6, grade for which the task was originally designed; (NUMBER), number of items in the cluster.

In Grades 1–3, instructions were provided in written form, on-screen, and with a pre-recorded voiceover to avoid any reading difficulties and to ensure greater validity of the assessments. Thus, students used headphones during the administration of the tests. After listening to the instructions, they indicated their answer by using the mouse or keyboard (in the case of desktop computers, which are most commonly used in Hungarian schools) or by directly tapping, typing or dragging the elements of the tasks using their fingers on tablets.

The tasks presented in **Figure 1** assess students' mathematical and scientific reasoning. Based on the framework for the diagnostic assessment of mathematics (Csapó and Szendrei, 2011) and science (Csapó and Szabó, 2012), the main questions in this psychological dimension related to how well mathematics and science education was adjusted to students' psychological development, how learning mathematics and science could contribute to the development of specific reasoning skills and how effectively they could stimulate students' general cognitive development. Items developed to measure the psychological dimension of learning encompassed a long list of skills, such as inductive reasoning, deductive reasoning, analogical reasoning, combinatorial reasoning, systematization skills, proportional reasoning and correlative reasoning. Two examples of tasks for assessing students' inductive reasoning are presented in **Figure 1**. Students had to discover regularities by detecting dissimilarities with respect to attributes of different objects. They completed the tasks by dragging the elements to different areas, thereby defining the proper sets. The scoring of all tasks was automated, including items with several correct answers.

Figure 2 presents a task measuring student's science disciplinary knowledge and a mathematics tasks measuring the application dimension. In the science task, students retrieve disciplinary knowledge of phases of the water cycle. In the mathematics task, students have to select and place flowers – drag and drop – in the vase; only the number of flowers counts. The task measures the application of adding up to 10 in a realistic application context.

Procedures

The tests were administered over a period of 7 weeks in computer rooms within the participating schools during regular school hours. Each test lasted approximately 45 min. Test sessions were supervised by teachers who had been thoroughly trained in test administration. The tests were delivered on the eDia online platform. After students entered the system and chose the domain (reading, mathematics, or science), the system randomly selected a test for that student from the nine tests available in the appropriate grade level.

To learn to use the program, students were provided with instructions and a trial (warm-up) task with immediate feedback. This instruction included: (1) a yellow bar at the top of the screen to show how far along they were on the test; (2) they had to click on the speaker icon to listen to the task instructions; (3) they had to click on the "next" button to move on to the next task; (4) pupils in Grades 1 and 2 received extra warm-up tasks to enhance keyboarding and mouse skills; and (5) after completing the last task, participants received immediate visual feedback with a display of 1 to 10 balloons, where the number of balloons was proportionate to their achievement.

The feedback system available for the teacher was more elaborate. Due to the large number of students and items, the Rasch analyses were run with the built-in analytic module in the eDia system. As the tasks in the item bank were scaled using IRT, it was possible to compare students' achievement. Teachers received feedback on students' achievement both as a percentage of correct items and as ability scores. For each grade and domain, the national average achievement (ability score) was set at 500 with a standard deviation of 100 (Carlson, 2009; Ferrão et al., 2015; Weeks, 2018). This was the point of reference for interpreting students' achievement.

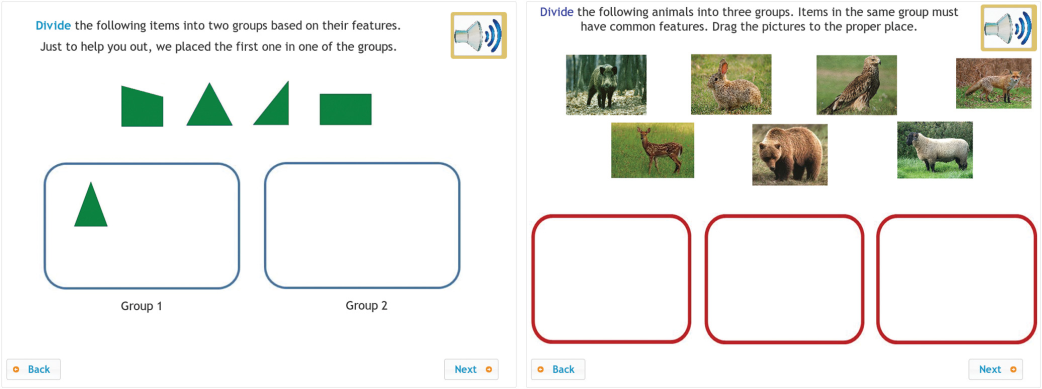
We used confirmatory factor analyses (CFA) within structural equation modeling (SEM) (Bollen, 1989) to test the underlying measurement models of reading, mathematics and science knowledge in the three dimensions of learning: psychological (reasoning), application (literacy), and disciplinary knowledge, respectively (RQ 1). We used the preferred estimator for categorical variables; the adjusted weighted least squares mean and variance (WLSMV) (Muthén and Muthén, 2012). We tested a 3-dimensional model to distinguish the three different dimensions of learning, and we also tested a 1-dimensional model with all three dimensions combined under one general factor. In order to test which model fitted the data better, we carried out a special χ^2-difference test in Mplus. We also used CFA to test the underlying measurement model, and to determine the invariance behavior of the psychological dimension across the three domains of learning (RQ 2).

To establish a developmentally valid scale, we used the Rasch model with the vertical and horizontal scaling of the data (RQs 2 and 4) and then a linear transformation of the logit metric. As indicated above, for each domain and at each grade level, the mean achievement of each dimension was set to 500 with a standard deviation of 100. We used path models to test the effect and predictive power of school learning on the psychological dimension of learning (RQ 3).

RESULTS

The Psychological Dimension of Learning

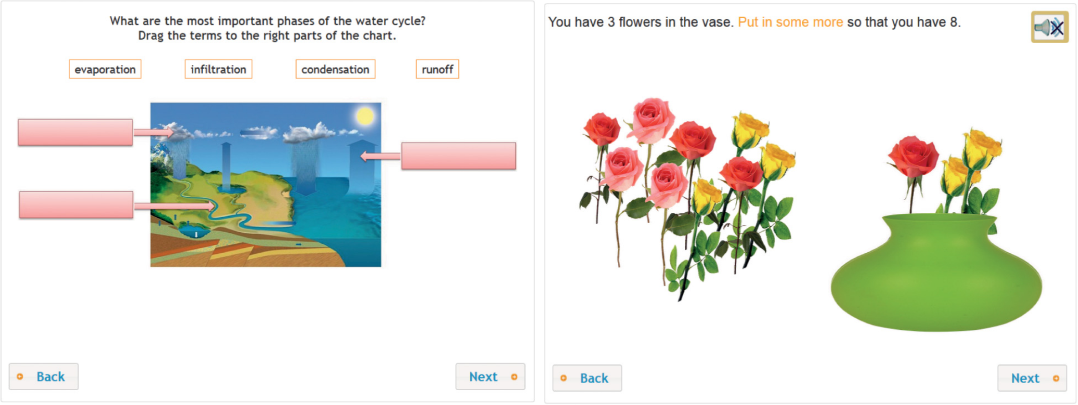
Results showed that the psychological (reasoning/thinking), application and disciplinary dimensions of learning can be distinguished empirically and are independent of domain and grade. The χ^2-difference test in Mplus showed that the 3-dimensional model fitted significantly better than the 1-dimensional model in each grade and in each domain (see **Tables 3–5** for reading, mathematics and science, respectively). Generally, the 3-dimensional measurement model for each domain showed a good model fit (**Tables 3–5**), based on Hu and Bentler's (1999) recommended cut-off values. The comparative fit index (CFI) and the Tucker–Lewis index (TLI) values above 0.95 and the root mean square error of approximation (RMSEA) below 0.06 indicated a good global model fit.



Divide the following items into two groups based on their features. Just to help you out, we placed the first one in one of the groups.

Divide the following animals into three groups. Items in the same group must have common features. Drag the pictures to the proper place.

FIGURE 1 | Measuring the psychological dimension of learning: assessment of students' inductive reasoning skills in the context of geometry and biology.



What are the most important phases of the water cycle? Drag the terms to the right parts of the chart.

You have 3 flowers in the vase. Put in some more so that you have 8.

FIGURE 2 | Measuring the disciplinary dimension of learning science and the application dimension of learning mathematics.

TABLE 3 | Goodness of fit indices for testing the dimensionality of reading from Grades 1 to 8.

Grade	Model	χ^2	df	p	$\Delta\chi^2$	Δdf	p	CFI	TLI	RMSEA	90% C.I.
1	3-dim.	378.262	296	0.001	29.055	3	0.001	0.947	0.941	0.057	[0.038, 0.073]
	1-dim.	448.137	299	0.001				0.903	0.895	0.076	[0.061, 0.090]
2	3-dim.	514.018	461	0.001	30.963	3	0.001	0.975	0.973	0.032	[0.026, 0.075]
	1-dim.	575.196	464	0.001				0.948	0.945	0.047	[0.033, 0.059]
3	3-dim.	406.497	347	0.01	15.681	3	0.01	0.833	0.818	0.054	[0.026, 0.075]
	1-dim.	430.585	350	0.01				0.773	0.755	0.062	[0.039, 0.082]
4	3-dim.	592.821	431	0.01	90.820	3	0.001	0.937	0.932	0.066	[0.052, 0.079]
	1-dim.	695.499	434	0.01				0.898	0.891	0.084	[0.072, 0.095]
5	3-dim.	2046.006	125	0.001	92.737	3	0.001	0.911	0.908	0.035	[0.027, 0.041]
	1-dim.	2276.042	122	0.001				0.839	0.833	0.046	[0.041, 0.052]
6	3-dim.	530.220	431	0.001	77.918	3	0.001	0.970	0.967	0.037	[0.025, 0.047]
	1-dim.	755.989	434	0.001				0.902	0.895	0.066	[0.058, 0.073]
7	3-dim.	1078.340	899	0.001	110.370	3	0.001	0.969	0.967	0.030	[0.022, 0.036]
	1-dim.	1458.058	902	0.001				0.904	0.899	0.052	[0.047, 0.057]
8	3-dim.	696.816	524	0.001	76.199	3	0.001	0.974	0.972	0.035	[0.028, 0.042]
	1-dim.	979.228	527	0.001				0.933	0.928	0.057	[0.052, 0.063]

df, degrees of freedom; CFI, comparative fit index; TLI, Tucker–Lewis index; RMSEA, root mean square error of approximation; χ^2 and df were estimated by WLSMV. $\Delta\chi^2$ was estimated with the difference test procedure in MPlus (see Muthén and Muthén, 2012). C.I., confidence interval.

TABLE 4 | Goodness of fit indices for testing the dimensionality of mathematics from Grades 1 to 8.

Grade	Model	χ^2	df	p	$\Delta\chi^2$	Δdf	p	CFI	TLI	RMSEA	90% C.I.
1	3-dim.	409.506	249	0.001	95.309	3	0.001	0.953	0.948	0.077	[0.063, 0.090]
	1-dim.	586.328	252	0.001				0.902	0.893	0.110	[0.099, 0.122]
2	3-dim.	543.407	321	0.001	96.826	3	0.001	0.944	0.939	0.061	[0.052, 0.070]
	1-dim.	734.133	324	0.001				0.897	0.889	0.083	[0.075, 0.091]
3	3-dim.	171.573	149	0.01	15.784	3	0.01	0.923	0.912	0.046	[0.000, 0.075]
	1-dim.	194.581	152	0.01				0.855	0.837	0.063	[0.032, 0.087]
4	3-dim.	236.477	206	0.01	40.265	3	0.001	0.940	0.933	0.060	[0.000, 0.093]
	1-dim.	268.352	209	0.01				0.883	0.871	0.083	[0.050, 0.111]
5	3-dim.	381.365	186	0.001	110.584	3	0.001	0.939	0.931	0.060	[0.052, 0.069]
	1-dim.	675.939	189	0.001				0.847	0.830	0.095	[0.087, 0.102]
6	3-dim.	680.214	492	0.001	112.972	3	0.001	0.912	0.906	0.054	[0.043, 0.063]
	1-dim.	966.684	495	0.001				0.780	0.765	0.085	[0.077, 0.093]
7	3-dim.	1182.063	816	0.001	205.034	3	0.001	0.968	0.966	0.047	[0.041, 0.052]
	1-dim.	1882.948	819	0.001				0.908	0.903	0.079	[0.075, 0.084]
8	3-dim.	3021.062	557	0.001	165.118	3	0.001	0.876	0.867	0.124	[0.120, 0.128]
	1-dim.	3412.642	560	0.001				0.856	0.847	0.133	[0.129, 0.137]

df, degrees of freedom; CFI, comparative fit index; TLI, Tucker–Lewis index; RMSEA, root mean square error of approximation; χ^2 and df were estimated by WLSMV. $\Delta\chi^2$ was estimated with the difference test procedure in MPlus (see Muthén and Muthén, 2012). C.I., confidence interval.

TABLE 5 | Goodness of fit indices for testing the dimensionality of science from Grades 1 to 8.

Grade	Model	χ^2	df	p	$\Delta\chi^2$	Δdf	p	CFI	TLI	RMSEA	90% C.I.
1	3-dim.	596.485	461	0.001	57.623	3	0.001	0.921	0.915	0.050	[0.038, 0.061]
	1-dim.	659.870	464	0.001				0.886	0.878	0.060	[0.049, 0.073]
2	3-dim.	464.075	321	0.001	39.177	3	0.001	0.944	0.939	0.038	[0.030, 0.045]
	1-dim.	554.254	324	0.001				0.910	0.903	0.048	[0.041, 0.055]
3	3-dim.	732.349	431	0.01	66.500	3	0.01	0.924	0.918	0.111	[0.097, 0.124]
	1-dim.	786.319	434	0.01				0.911	0.904	0.119	[0.106, 0.133]
4	3-dim.	159.502	132	0.01	19.191	3	0.001	0.939	0.930	0.060	[0.000, 0.091]
	1-dim.	178.564	135	0.01				0.904	0.891	0.075	[0.041, 0.103]
5	3-dim.	571.944	402	0.001	151.940	3	0.001	0.938	0.933	0.040	[0.033, 0.048]
	1-dim.	950.437	405	0.001				0.801	0.787	0.072	[0.066, 0.078]
6	3-dim.	716.173	402	0.001	332.375	3	0.001	0.934	0.928	0.048	[0.063, 0.074]
	1-dim.	1925.098	405	0.001				0.679	0.655	0.106	[0.101, 0.111]
7	3-dim.	999.868	524	0.001	185.888	3	0.001	0.882	0.874	0.039	[0.035, 0.042]
	1-dim.	1564.230	527	0.001				0.743	0.726	0.057	[0.054, 0.060]
8	3-dim.	664.189	374	0.001	112.367	3	0.001	0.882	0.872	0.041	[0.036, 0.046]
	1-dim.	897.133	377	0.001				0.788	0.772	0.055	[0.050, 0.060]

df, degrees of freedom; CFI, comparative fit index; TLI, Tucker–Lewis index; RMSEA, root mean square error of approximation; χ^2 and df were estimated by WLSMV. $\Delta\chi^2$ was estimated with the difference test procedure in MPlus (see Muthén and Muthén, 2012). C.I., confidence interval.

In most cases, the 3-dimensional models fitted the data significantly better than that the 1-dimensional models. In some cases, mostly in Grades 7 and 8, the 3-dimensional model fit indices were lower. This could have been because the tasks were originally developed for students in lower grades.

The fit indices dropped in the case of mathematics and science in Grade 8 but were significantly higher than that of the 1-dimensional model. Thus, the psychological, application and disciplinary dimensions of learning could be distinguished. The psychological dimension of learning could be made visible independently of the measured domain in everyday educational settings, thus supporting Hypothesis 1.

The Psychological Dimension of Learning Across Domains

The bivariate correlations of the psychological dimensions between pairs of domains (mathematics and reading, mathematics and science, and reading and science) ranged from 0.29 to 0.49 and were statistically significant (**Table 6**). At each grade level, the correlations of the psychological dimension (reasoning/thinking) tended to be the highest between mathematics and reading and lowest between mathematics and science. The strongest set of correlations, independent of the measured domain, was found in Grade 8, indicating that the

TABLE 6 | Correlations of the psychological dimension between pairs of domains from Grades 1 to 8.

	Correlations of the psychological dimension		
Grade	**Between mathematics and reading**	**Between mathematics and science**	**Between reading and science**
1	0.426	0.372	0.407
2	0.435	0.342	0.421
3	0.390	0.289	0.340
4	0.452	0.421	0.420
5	0.436	0.404	0.429
6	0.437	0.421	0.398
7	0.440	0.429	0.395
8	0.493	0.452	0.438

All coefficients are significant at $p < 0.0001$ level.

psychological dimension of learning in reading, mathematics and science were highly correlated, but not identical constructs.

The invariance in the psychological dimension of learning across the three domains was supported by comparing the 3-dimensional measurement model, which distinguishes the psychological dimension of reading, mathematics and science, and the 1-dimensional measurement model, which combines the psychological dimension of the different learning domains under a single factor. The special χ^2-difference test in Mplus showed that the 3-dimensional model fitted significantly better at each grade level than the 1-dimensional model (**Table 7**).

The Rate of Development in the Psychological Dimension

Figure 3 presents the mean cognitive development scale scores in the psychological dimension of learning reading, mathematics and science. Please note that in each domain, the mean score of Grade 8 students was set at 500 with a standard deviation of 100, thereby constructing the point of reference for interpreting students' achievement. This means that we cannot compare the development of the psychological dimension of learning across domains, but we can compare the rate of development.

We found that the amount and rate of cognitive development were almost the same in each domain between Grades 6 and 8 and that there was no appreciable development in reading and science between Grades 2 and 6. The greatest rate of progress occurred in Grade 1 in reading and science, but not mathematics. Generally, there was a steady increase in the psychological dimension of learning in mathematics, especially in the first 4 years of schooling. The results confirmed our hypothesis that children's cognitive development is slow (Molnár et al., 2013; Molnár et al., 2017), thus indicating the importance of the explicit development in this dimension in school lessons. Overall, these results highlighted the importance, sensitivity and potential of the development of thinking skills in the early years of schooling.

Relationship Between the Three Dimension of Learning

The possibility and practical relevance of separating the psychological dimension of learning can be explored from another perspective by examining the proportion of its variance that remains unexplained if the more readily visible disciplinary and application dimensions (referred to together as school knowledge) are taken into account. Technically, these dimensions may be considered as potential predictors of the psychological dimension.

We used continuous factor indicators in SEM analyses to examine the relationships between school knowledge and the

TABLE 7 | Goodness of fit indices for testing the dimensionality of the psychological dimension in reading, mathematics, and science using 1- and 3-dimensional models for Grades 1 to 8.

Grade	Model	χ^2	df	p	$\Delta\chi^2$	Δdf	p	CFI	TLI	RMSEA	90% C.I.
1	3-dim.	406.929	321	0.001	55.558	3	0.001	0.963	0.959	0.046	[0.031, 0.059]
	1-dim.	510.320	324	0.001				0.919	0.912	0.067	[0.056, 0.078]
2	3-dim.	282.857	167	0.001	75.885	3	0.001	0.890	0.903	0.062	[0.049, 0.074]
	1-dim.	449.393	170	0.001				0.765	0.738	0.095	[0.085, 0.106]
3	3-dim.	180.800	167	0.001	20.178	3	0.001	0.921	0.910	0.036	[0.000, 0.069]
	1-dim.	203.772	170	0.001				0.806	0.783	0.056	[0.014, 0.082]
4	3-dim.	209.681	206	0.001	42.211	3	0.001	0.990	0.989	0.018	[0.000, 0.060]
	1-dim.	289.741	209	0.001				0.775	0.751	0.083	[0.058, 0.105]
5	3-dim.	398.477	296	0.001	126.509	3	0.001	0.934	0.928	0.039	[0.028, 0.049]
	1-dim.	755.052	299	0.001				0.707	0.681	0.082	[0.075, 0.089]
6	3-dim.	592.088	431	0.001	80.817	3	0.001	0.901	0.890	0.078	[0.062, 0.093]
	1-dim.	785.255	434	0.001				0.767	0.750	0.115	[0.102, 0.128]
7	3-dim.	1154.972	699	0.001	187.282	3	0.001	0.912	0.906	0.059	[0.053, 0.065]
	1-dim.	1893.066	702	0.001				0.769	0.757	0.095	[0.090, 0.100]
8	3-dim.	471.630	347	0.001	142.432	3	0.001	0.918	0.911	0.042	[0.031, 0.059]
	1-dim.	747.482	350	0.001				0.740	0.719	0.072	[0.065, 0.079]

df, degrees of freedom; CFI, comparative fit index; TLI, Tucker–Lewis index; RMSEA, root mean square error of approximation; χ^2 and df were estimated by WLSMV. $\Delta\chi^2$ was estimated with the difference test procedure in MPlus (see Muthén and Muthén, 2012). C.I., confidence interval.

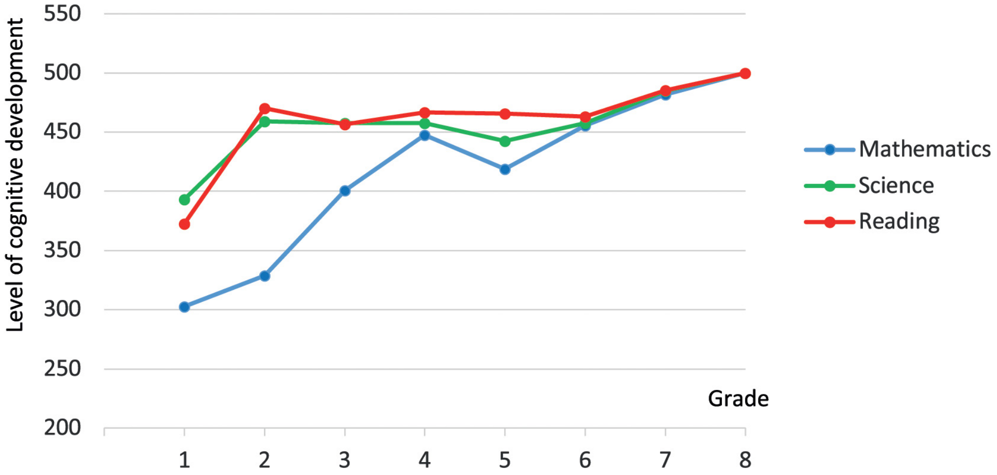


FIGURE 3 | The speed of the cognitive development in the psychological dimension of learning within the domains of mathematics, science and reading (Please note, that in each measured domain the mean of the 8th graders' achievement was artificially set to 500 with a standard deviation of 100).

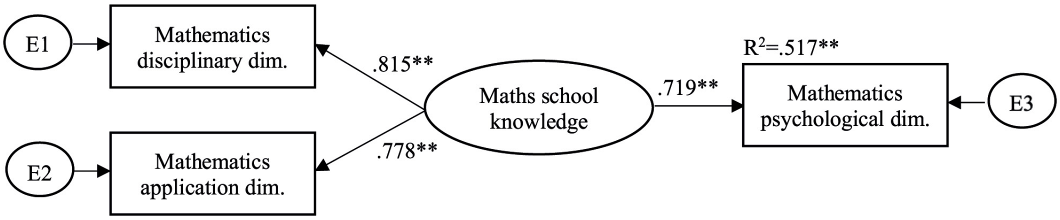

E1
Mathematics disciplinary dim.
E2
Mathematics application dim.
.815**
.778**
Maths school knowledge
.719**
R²=.517**
Mathematics psychological dim.
E3

FIGURE 4 | A structural model of mathematics school knowledge as a predictor of students' cognitive development in the domain of mathematical reasoning.

psychological dimension of learning in each domain. School knowledge as a latent factor was specified as the application and disciplinary dimensions of learning. According to the results, school knowledge predicted the psychological dimension of learning in all domains, but a significant amount of variance remained unexplained (see **Figures 4–6**). This indicates that existing aspects of the psychological dimension of learning can be separated from school knowledge as measured by the disciplinary and application parts of students' knowledge. That is, it is relevant to measure the psychological dimension of learning in addition to measuring the disciplinary and application dimensions of learning. So our hypothesis was confirmed.

The amount of explained variance was statistically significant and almost the same for mathematics and reading and somewhat higher for science. This suggests that there may be more common reasoning aspects in the three dimensions of science. The model for each domain fitted well (CFI = 1.000, TLI = 1.000, RMSEA = 0.000).

Gender Difference in the Psychological Dimension of Learning

In the present study, girls outperformed boys in the psychological dimension of learning in reading, mathematics and science (Mathematics: $F = 0.272$, $t = -6.696$, $p < 0.001$; Science: $F = 3.578$, $t = -11.525$, $p < 0.001$; Reading: $F = 3.224$, $t = -4{,}370$, $p < 0.001$); however, this varied by grade level (see **Table 8**). The largest, statistically significant differences in favor of girls were found in Grades 4 and 5, where girls outperformed boys in all three domains, and in Grades 6 to 8, where girls outperformed boys in two of the three domains. Girls also outperformed boys in reading in Grades 3–8, in mathematics in Grades 1 and 4–6, and in science in Grades 1, 4, 5, 7, and 8.

In this section, we examine gender differences among Grade 8 students – the grade level of students in PISA, TIMSS and our study. The results confirm our hypotheses that an assessment which focuses on students' disciplinary knowledge or application does not replace an assessment of the psychological dimension of learning. In the case of mathematics, no gender differences were detected in the application and psychological dimensions of learning, but girls scored significantly higher, on average, than boys in the disciplinary dimension of learning. The results were different in the case of science. There were no gender differences in the application dimension of science learning. Boys achieved significantly higher in the psychological dimension.

DISCUSSION

Previous research has already identified several characteristics of learning reading, mathematics and science. However, it has mainly focussed on only one dimension; either the disciplinary

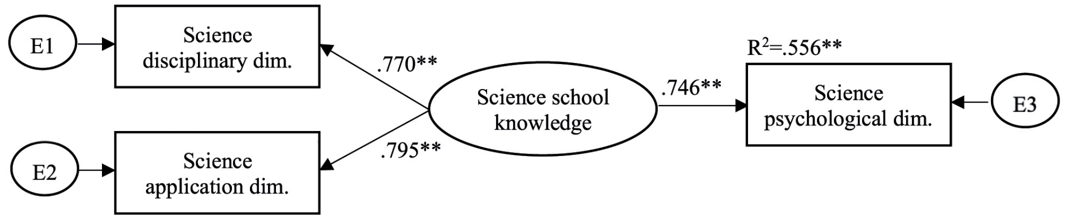


FIGURE 5 | A structural model of science school knowledge as a predictor of students' cognitive development in the domain of scientific reasoning.

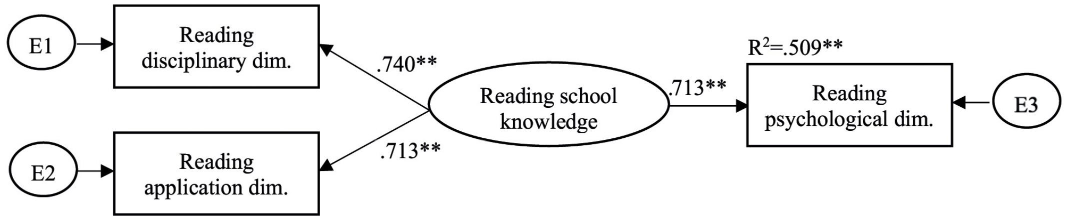


FIGURE 6 | A structural model of reading school knowledge as a predictor of students' cognitive development in the domain of the psychological dimension of reading.

TABLE 8 | Gender differences in the psychological dimension of learning in reading, mathematics and science in Grades 1 to 8.

Grade	Area	N	Boys' mean (SD)	Girls' mean (SD)	F	p	t	p	d
1	R	685	346 (128)	359 (123)	0.075	0.784	−1.404	0.161	0.107
	M	707	306 (117)	332 (129)	2.058	0.152	−2.871	0.004	0.215
	S	487	371 (98)	392 (121)	7.794	0.005	−2.143	0.033	0.030
2	R	1024	457 (113)	462 (118)	0.387	0.534	−0.705	0.481	0.044
	M	1033	338 (137)	338 (144)	0.252	0.616	−0.031	0.975	0.000
	S	668	444 (100)	456 (102)	0.258	0.612	−1.505	0.133	0.118
3	R	1152	439 (126)	463 (123)	0.736	0.391	−3.214	0.001	0.192
	M	1236	420 (103)	417 (109)	2.389	0.122	0.465	0.642	0.028
	S	829	455 (89)	460 (93)	0.227	0.634	−0.728	0.467	0.054
4	R	1539	452 (107)	473 (106)	0.237	0.627	−3.839	0.000	0.197
	M	1567	451 (99)	465 (101)	0.638	0.425	−2.832	0.005	0.139
	S	862	443 (109)	463 (100)	2.178	0.140	−2.772	0.006	0.191
5	R	1721	447 (113)	479 (104)	3.374	0.066	−6.209	0.000	0.294
	M	1877	422 (95)	439 (100)	1.849	0.174	−3.798	0.000	0.174
	S	1540	429 (104)	453 (96)	4.063	0.044	−4.725	0.000	0.239
6	R	1559	445 (105)	480 (99)	1.466	0.226	−6.858	0.000	0.342
	M	1496	460 (89)	469 (96)	1.991	0.158	−2.035	0.042	0.097
	S	1469	452 (111)	457 (109)	0.113	0.737	−0.842	0.400	0.045
7	R	1280	465 (107)	490 (97)	2.239	0.135	−4.341	0.000	0.244
	M	1291	481 (99)	491 (88)	4.528	0.034	−1.859	0.063	0.106
	S	1250	480 (98)	493 (96)	0.100	0.752	−2.259	0.024	0.134
8	R	1035	481 (101)	515 (96)	3.165	0.076	−5.429	0.000	0.345
	M	932	494 (102)	505 (96)	1.703	0.192	−1.749	0.081	0.111
	S	954	490 (99)	509 (98)	0.014	0.906	−2.958	0.003	0.192

R, reading; M, mathematics; S, science; F, F-value; t, t-value; p, significance level; d, Cohen-d.

dimension or the application dimension of learning, and on the reading, mathematics and science learning of older students. There have been significant attempts to concentrate on the application and reasoning dimensions, but educational practice has mostly focussed on the assessment of the content of individual curriculum subjects. The application of knowledge has seldom been assessed, although the PISA assessments have highlighted its importance. Because of the lack of easy-to-use assessments, the psychological dimension of learning (cognitive development and reasoning) remains hidden. Therefore, neither the students nor their teachers receive feedback on level or development in this dimension. This study provides evidence that the psychological dimension of learning can be made visible and that technology-based assessment may be applied in an everyday educational context. This evidence highlights the importance of the assessment and the explicit development of the psychological dimension of learning in a school context. Further, it points to gender differences in the developmental rate of the psychological dimension of learning in favor of girls, although this varies by grade and domain.

Results support our hypotheses that the three dimensions of learning can be distinguished empirically and can be assessed separately. The 3-dimensional frameworks derived from previous research, including international comparative studies (Csapó and Szendrei, 2011; Csapó and Csépe, 2012; Csapó and Szabó, 2012), showed relatively good validity, and the results from the current analyses confirmed that they may form evidence-based foundations for diagnostic assessment. The most important findings from these analyses was that the psychological dimension of learning can be measured at the primary school level in the context of three of the most important domains of learning – reading, mathematics and science.

The present results also confirmed that, although the roots of the psychological development of different domains are universal and the domains of learning build on each other (Molnár and Csapó, 2019), there are still significant developmental differences between them. While there is a close connection between the development of early literacy and numeracy, and later mathematics learning builds on reading, and science builds on both (McKeachie, 1987), our results support the notion that the transfer is not obvious between the different domain contexts. There were statistically significant correlations between the development scores in the psychological dimension of reading, mathematics and science learning, but they were not identical constructs.

Previous studies have indicated that children's cognitive development is slow (Molnár et al., 2013, 2017) but that it can be taught effectively (de Koning et al., 2002; Klauer and Phye, 2008; Perret, 2015). Our results confirmed both of these notions as there was no appreciable development in the psychological dimension of learning in reading and science for students in Grades 2–6, and students' cognitive development was the most steady (and effective) in mathematics, where the greatest development took place in the first years of schooling. This confirms previous research findings and highlights the potential of developing thinking skills in the early years of schooling.

The results of the SEM indicated the complex nature of learning in reading, mathematics and science. An examination of the predictive power of school knowledge on the psychological dimension of learning showed that the disciplinary and application dimensions of learning together predicted the psychological dimension of learning at a moderate, but statistically significant level, while a significant amount of variance remained unexplained. This indicates that school knowledge in reading, mathematics, and science can contribute to the development of the psychological dimension of learning and can stimulate students' general cognitive development, but the transfer effect may not be high. The results suggest that aspects of the psychological dimension of learning exist and can be separated from the learning dimensions assessed most often at school and in international comparative studies. This highlights the importance and relevance of developing measures of the psychological dimension of learning as well.

To provide context to interpret the size of the gender difference in the psychological dimension of learning in reading, mathematics and science, we compared our results to findings on gender differences in the most prominent international comparative studies. The gender differences in the international studies at Grade 4 and 8 were found in our study. We found gender differences in reading over almost all the primary school grade levels, including Grades 4 and 8, indicating that girls perform better in reading, irrespective of the dimension of learning.

LIMITATIONS OF THE PRESENT STUDY

As the PISA 2015, TIMSS 2015, and PIRLS 2016 studies have also indicated, there are large differences between countries not only on the level of reading, mathematics and science performance, but also in gender differences. Therefore, results found in one country cannot be generalized across countries and cultures. Although general trends have been found, the generalizability of the results may be limited. The method we applied in this study was generalizable and may be useful for making the psychological dimension of learning visible in any educational context. A further limitation of the study could be the results of the "common method bias" and "test motivation" as possible sources of shared variance across tests and domains. Participation in the study was voluntary, and although the large sample sizes and the diversity of the schools made the results sufficiently robust, the actual samples were not nationally representative. Thus, the present study does not provide a complete picture of the Hungarian education system. Nevertheless, the analyses did reveal some generalizable trends.

CONCLUSION

The 3-dimensional frameworks for the diagnostic assessment used in the present study were devised on the basis of current results from a number of research fields ranging from cognitive neuroscience to research on cognitive development, standard setting and the theoretical frameworks of large-scale international comparative studies. The item banks for assessing reading, mathematics and science were developed through the careful mapping of assessment tasks onto frameworks. The next step in scientifically establishing and further developing the diagnostic system is to empirically validate the 3-dimensional framework. We first presented the results of the comprehensive analyses in this study. In the present analyses, we focused on the psychological dimension of learning, which determines the dimensions of disciplinary knowledge and application, but is less visible or observable in the school context.

The results confirmed the theoretical foundations of the project and made clear that the psychological dimension can be distinguished and measured in the context of the most important domains of learning in the beginning phase of schooling. These findings indicate directions for further research as well. Item development for this study was based on the theoretical frameworks without empirical evidence of dimensionality. Based on the empirical confirmation of the three dimensions in this study, the validity of the assessment scales constructed from the item banks, may be improved by exploring how well the items fit particular scales.

Establishing scales empirically to assess the psychological dimension of learning paves the way to improving learning as well. The evidence that cognitive development is measurable provides a basis for large-scale systematic diagnostic monitoring of the development of students' thinking skills, one of the most sorely lacking elements in the current spectrum of assessment practices. It also supports different types of intervention studies from teacher-initiated practical improvements to well-controlled, randomized experiments.

ETHICS STATEMENT

The authors only had access to anonymized data, and hence an ethics approval and parental consent were not required as per applicable institutional and national guidelines and regulations. The assessment data collected for this study formed integrated parts of the normal educational processes of the participating schools. The coding system for the online platform masked students' identity, the researchers would thus have been unable to tie the data to the students. The results from the low-stakes diagnostic assessments were only disclosed to the participating students (as immediate feedback) and to their teachers. Because of the anonymity and low-stakes testing design of the assessment process, it was not required or possible to request and obtain written informed parental consent from the participants.

AUTHOR CONTRIBUTIONS

GM and BC took responsibility for the content, including participation in the concept, design, analysis, drafting the manuscript, writing and final approval of the manuscript, and agreed to be accountable for all aspects of the study.

FUNDING

This study was funded by OTKA K115497 and EFOP 3.2.15.

ACKNOWLEDGMENTS

We would like to express our appreciation to the members of the eDia Research Group, who provided continuous support for the participating schools during data collection. A special thanks go to Dr. Géza Makay and Dóra Mokri, who assisted us with data filtering and scaling. We also wish to thank the two the reviewers for their valuable comments on earlier versions of this manuscript.

REFERENCES

Adey, P. (2007). "The CASE for a general factor in intelligence," in *Integrating the Mind: Domain General Versus Domain Specific Processes in Higher Cognition*, ed. M. J. Roberts (Hove: Psychology Press), 369–385.

Adey, P., and Csapó, B. (2012). "Developing and assessing scientific reasoning," in *Framework for Diagnostic Assessment of Science*, eds B. Csapó and G. Szabó (Budapest: Nemzeti Tankönyvkiadó), 17–53.

Adey, P., Csapó, B., Demetriou, A., Hautamaki, J., and Shayer, M. (2007). Can we be intelligent about intelligence? Why education needs the concept of plastic general ability. *Educ. Res. Rev.* 2, 75–97.

Adey, P., and Shayer, M. (1994). *Really Raising Standards: Cognitive Intervention and Academic Achievement*. London: Routledge.

Adey, P., Shayer, M., and Yates, C. (2001). *Thinking Science: The Curriculum Materials of the CASE Project*, 3rd Edn. London: Nelson Thornes.

Alexander, P. A., and Murphy, P. K. (1999). Nurturing the seeds of transfer: a domain-specific perspective. *Int. J. Educ. Res.* 31, 561–576. doi: 10.1016/s0883-0355(99)00024-5

Binet, A., and Simon, T. (1916). "The development of intelligence in children: The Binet–Simon Scale," in *The Development of Intelligence in Children*, eds E. S. Kite Trans. (Vineland NJ: Publications of the Training School).

Blomert, L., and Csépe, V. (2012). "Psychological foundations of reading acquisition and assessment," in *Framework for Diagnostic Assessment of Reading*, eds B. Csapó1 and V. Csépe (Budapest: Nemzeti Tankönyvkiadó), 17–77.

Bloom, B. S. (1969). *Cross-National Study of Educational Attainment: Stage I of the I.E.A. Investigation in Six Subject Areas*, Vol. 1–2. Washington, DC: Office of Education (DHEW).

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: Wiley.

Bransford, J. D., and Schwartz, D. L. (1999). Rethinking transfer: a simple proposal with multiple implications. *Rev. Res. Educ.* 24, 61–100. doi: 10.3102/0091732x024001061

Burstein, L. (1993). *The IEA Study. of Mathematics III: Student Growth and Classroom Processes*. Oxford: Pergamon Press.

Carlson, J. E. (2009). "Statistical models for vertical linking," in *Statistical Models for Test Equating, Scaling, and Linking*, ed. A. A. von Davier (New York, NY: Springer), 59–70. doi: 10.1007/978-0-387-98138-3_4

Carroll, J. B. (1993). *Human Cognitive Abilities*. Cambridge: Cambridge University Press.

Costa, A., and Kallick, B. (2009). *Learning and Leading with Habits of Mind: 16 Characteristics of Success*. Alexandria, VA: Association for Supervision and Curriculum Development.

Csapó, B. (1997). The development of inductive reasoning: cross-sectional assessments in an educational context. *Int. J. Behav. Dev.* 20, 609–626. doi: 10.1080/016502597385081

Csapó, B. (2010). Goals of learning and the organization of knowledge. *Zeitschrift für Pädagogik* 2010(Suppl. 56), 12–27.

Csapó, B., and Csépe, V. (2012). *Framework for Diagnostic Assessment of Reading*. Budapest: Nemzeti Tankönyvkiadó.

Csapó, B., Csíkos, C., and Molnár, G. (2015a). *A Matematikai Tudás Online Diagnosztikus Értékelésének Tartalmi Keretei [Framework for Online Diagnostic Assessment of Mathematics]*. Budapest: Nemzeti Tankönyvkiadó.

Csapó, B., Korom, E., and Molnár, G. (2015b). *A Természettudományi Tudás Online Diagnosztikus Értékelésének Tartalmi Keretei [Framework for Online Diagnostic Assessment of science]*. Budapest: Nemzeti Tankönyvkiadó.

Csapó, B., Steklács, J., and Molnár, G. (2015c). *Az Olvasás-Szövegértés Online Diagnosztikus Értékelésének Tartalmi Keretei [Framework for Online Diagnostic Assessment of Reading]*. Budapest: Nemzeti Tankönyvkiadó.

Csapó, B., and Szabó, G. (2012). *Framework for Diagnostic Assessment of Science*. Budapest: Nemzeti Tankönyvkiadó.

Csapó, B., and Szendrei, M. (2011). *Framework for Diagnostic Assessment of Mathematics*. Budapest: Nemzeti Tankönyvkiadó.

de Koning, E. (2000). *Inductive Reasoning in Primary Education: MEASUREMENT, Teaching, Transfer*. Zeist: Kerckebosch.

de Koning, E., Hamers, J. H. M., Sijtsma, K., and Vermeer, A. (2002). Teaching and transfer of inductive reasoning in primary education. *Dev. Rev.* 22, 211–241. doi: 10.1080/0361073X.2017.1398802

Demetriou, A., Efklides, A., Platsidou, M., and Campbell, R. L. (1993). The architecture and dynamics of a developing mind: experiential structuralism as a frame for unifying cognitive developmental theories. *Monogr. Soc. Res. Child Dev.* 58:167.

Demetriou, A., Gustafsson, J. E., Efklides, A., and Platsidou, M. (1992). "Structural systems in developing cognition, science, and education," in *Neo-Piagetian Theories of Cognitive Development: Implications and Applications for Education*, eds A. Demetriou, M. Shayer, and A. Efklides (London: Routledge), 79–103.

Ferrão, M. E., Costa, P. M., and Oliveira, P. N. (2015). Generalized partial credit item response model: linking scales in the assessment of learning. *J. Interdisciplinary Mathematics* 18, 339–354. doi: 10.1080/09720502.2014.932119

Feuerstein, R., Rand, Y., Hoffman, M., and Miller, M. (1980). *Instrumental Enrichment: An Intervention Program for Cognitive Modifiability*. Baltimore: University Park Press.

Geake, J., and Cooper, P. (2003). Cognitive neuroscience: implications for education? *Westmin. Stud. Educ.* 26, 7–20. doi: 10.1080/0140672032000070710

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., and Csapó, B. (2013). Complex problem solving in educational contexts – something beyond g: concept, assessment, measurement invariance, and construct validity. *J. Educ. Psychol.* 105, 364–379. doi: 10.1037/a0031856

Griffin, P., and Care, E. (2014). *Assessment and Teaching of 21st century skills: Methods and (approach)*. Dordrecht: Springer.

Grissmer, D. W. (2000). The continuing use and misuse of SAT scores. *Psychol. Public Policy Law* 6, 223–232. doi: 10.1037/1076-8971.6.1.223

Hattie, J. (2009). *Visible Learning: A Synthesis of 800 Meta-Analyses Relating to Achievement*. New York, NY: Routledge.

Hattie, J. (2012). *Visible Learning for Teachers: Maximizing Impact on Learning*. New York, NY: Routledge.

Hattie, J., and Anderman, E. M. (2013). "Introduction," in *International Guide to Student Achievement*, eds J. Hattie and E. M. Anderman (New York, NY: Routledge).

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Husén, T. (1967). *International. Study of Achievement in Mathematics: A Comparison of Twelve Countries*, Vol. 1–2. Stockholm: Almqvist & Wiksell.

IEA (1988). *Science Achievement in Seventeen Countries: A Preliminary Report*. Oxford: Pergamon Press.

Inhelder, B., and Piaget, J. (1958). *The Growth of Logical Thinking*. London: Routledge and Kegan Paul.

Klauer, K. J. (1989a). *Denktraining für Kinder I*. Göttingen: Hogrefe.

Klauer, K. J. (1989b). Teaching for analogical transfer as a means of improving problem solving, thinking and learning. *Instruct. Sci.* 18, 179–192. doi: 10.1007/bf00053358

Klauer, K. J. (1991). *Denktraining für Kinder II*. Göttingen: Hogrefe.

Klauer, K. J. (1993). *Denktraining für Jugendliche*. Göttingen: Hogrefe.

Klauer, K. J. (1997). "Training inductive reasoning: A developmental programme of higher-order cognitive skills," in *Teaching Thinking in Europe. Inventory of European programmes*, eds J. H. M. Hamers and M. T. Overtoom (Utrecht: Sardes), 77–81.

Klauer, K. J., and Phye, G. D. (2008). Inductive reasoning: a training approach. *Rev. Educ. Res.* 78, 85–123. doi: 10.3102/0034654307313402

MacGilchrist, B., Reed, J., and Myers, K. (2004). *The Intelligent School*. London: Sage.

Martin, M. O., Mullis, I. V. S., Foy, P., and Hooper, M. (2016). *TIMSS 2015 International Results in Science*. Chestnut Hill, MA: TIMSS & PIRLS International Study.

Mayrath, M. C., Clarke Midura, J., Robinson, D. H., and Schraw, G. (eds) (2012). *Technology Based Assessments for 21st Century Skills: Theoretical and Practical Implications from Modern Research*. Charlotte, NC: IAP.

McKeachie, W. J. (1987). Cognitive skills and their transfer: discussion. *Int. J. Educ. Res.* 11, 707–712. doi: 10.1016/0883-0355(87)90010-3

Molnár, G., and Csapó, B. (2011). Az 1–11 évfolyamot átfogó induktív gondolkodás kompetenciaskála készítése a valószínûségi tesztelmélet alkalmazásával. [Constructing inductive reasoning competency scales for years 1–11 using IRT models]. *Magyar Pedagógia* 111, 127–140.

Molnár, G., and Csapó, B. (2019). "Technology-based diagnostic assessments for identifying early learning difficulties in mathematics," in *International Handbook of Mathematical Learning Difficulties*, eds A. Fritz-Stratmann, P. Räsänen, and V. Haase (Heidelberg: Springer), 683–707. doi: 10.1007/978-3-319-97148-3_40

Molnár, G., Greiff, S., and Csapó, B. (2013). Inductive reasoning, domain-specific and complex problem solving: relations and development. *Think. Skills Creat.* 9, 35–45. doi: 10.1016/j.tsc.2013.03.002

Molnár, G., Greiff, S., Wüstenberg, S., and Fischer, A. (2017). "Empirical study of computer-based assessment of domain-general dynamic problem solving skills," in *The Nature of Problem Solving*, eds B. Csapó and J. Funke (Paris: OECD), 123–143.

Mullis, I. V. S., and Martin, M. O. (eds) (2015). *PIRLS 2016 Assessment Framework*, 2nd Edn. Boston, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., and Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics*. Boston, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., and Hooper, M. (2017). *PIRLS 2016 International Results in Reading*. Boston, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., and Eberber, E. (eds) (2005). *TIMSS 2007 Assessment Frameworks*. Boston, MA: TIMSS & PIRLS International Study Center, Boston College.

Muthén, L. K., and Muthén, B. O. (2012). *Mplus User's Guide*, 7th Edn. Los Angeles: Muthén & Muthén.

Nguyen, T., Watts, T. W., Duncan, G. J., Clements, D. H., Sarama, J. S., Wolfe, C., et al. (2016). Which preschool mathematics competencies are most predictive of fifth-grade achievement? *Early Child. Res. Q.* 36, 550–560. doi: 10.1016/j.ecresq.2016.02.003

Nunes, T., and Csapó, B. (2011). "Developing and assessing mathematical reasoning," in *Framework for Diagnostic Assessment of Mathematics*, eds B. Csapó and M. Szendrei (Budapest: Nemzeti Tankönyvkiadó), 17–56.

OECD (1999). *Measuring Student Knowledge and Skills: A New Framework for Assessment*. Paris: OECD.

OECD (2003). *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving*. Paris: OECD.

OECD (2016). *PISA 2015 Results: Excellence and Equity in Education*, Vol. I. Paris: OECD.

Perret, P. (2015). Children's inductive reasoning: developmental and educational perspectives. *J. Cogn. Educ. Psychol.* 14, 389–408. doi: 10.1891/1945-8959.14.3.389

Polya, G. (1981). *Mathematics Discovery: On Understanding, Learning, and Teaching Problem Solving*. New York, NY: John Wiley & Sons.

Ritchhart, R., Church, M., and Morrison, K. (2011). *Making Thinking Visible: How to Promote Engagement, Understanding, and Independence for all Learners*. San Francisco: Jossey-Bass.

Shayer, M. (1999). Cognitive acceleration through science education II: its effect and scope. *Int. J. Sci. Educ.* 21, 883–902. doi: 10.1080/095006999290345

Shayer, M., and Adey, P. (eds) (2002). *Learning Intelligence: Cognitive Acceleration Across the Curriculum From 5 to 15 years*. Milton Keynes: Open University Press.

Shayer, M., and Adhami, M. (2007). Fostering cognitive development through the context of mathematics: results of the CAME project. *Educ. Stud. Math.* 64, 265–291. doi: 10.1007/s10649-006-9037-1

Thomas, M. S., Ansari, D., and Knowland, V. C. (2019). Annual research review: educational neuroscience: progress and prospects. *J. Child Psychol. Psychiatry* 60, 477–492. doi: 10.1111/jcpp.12973

Trilling, B., and Fadel, C. (2009). *21st Century Skills: Learning for Life in our Times*. San Francisco: John Wiley & Sons.

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Weeks, J. P. (2018). An application of multidimensional vertical scaling. *Measure. Interdiscipl. Res. Perspect.* 16, 139–154. doi: 10.1080/15366367.2018.1502005

Wüstenberg, S., Greiff, S., Molnar, G., and Funke, J. (2014). Determinants of cross-national gender differences in complex problem solving competency. *Learn. Individ. Diff.* 29, 18–29. doi: 10.1016/j.lindif.2013.10.006

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 27 June 2019
doi: 10.3389/fpsyg.2019.01429

The Skilled, the Knowledgeable, and the Motivated: Investigating the Strategic Allocation of Time on Task in a Computer-Based Assessment

*Johannes Naumann**

Institute of Educational Research, University of Wuppertal, Wuppertal, Germany

OPEN ACCESS

Edited by:
Ronny Scherer,
University of Oslo, Norway

Reviewed by:
Marlit Annalena Lindner,
Christian-Albrechts-Universität zu Kiel,
Germany
Gaston Saux,
National Council for Scientific
and Technical Research (CONICET),
Argentina

****Correspondence:***
Johannes Naumann
j.naumann@uni-wuppertal.de

Specialty section:
This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Psychology

***Received:** 02 January 2019*
***Accepted:** 04 June 2019*
***Published:** 27 June 2019*

Citation:
Naumann J (2019) The Skilled,
the Knowledgeable, and
the Motivated: Investigating
the Strategic Allocation of Time on
Task in a Computer-Based
Assessment.
Front. Psychol. 10:1429.
doi: 10.3389/fpsyg.2019.01429

In large scale low stakes assessments, students usually choose their own speed at which to work on tasks. At the same time, previous research has shown that in hard tasks, the time students invest is a positive predictor of task performance. From this perspective, a relevant question is whether student dispositions other than the targeted skill might affect students' time on task behavior, thus potentially affecting their task performance and in turn their estimated skill in the target domain. Using PISA 2009 computer based assessment data, the present research investigated for the domain of reading digital text whether three variables that can be assumed to predict performance in digital reading tasks, comprehension skill, enjoyment of reading, and knowledge of reading strategies would also predict how much time students would devote to digital reading tasks, and in particular, whether they would adapt time on task to task difficulty. To address this question, two linear mixed models were estimated that predicted the time students spent on a task, and the average time students spent on relevant pages within each task, by the interaction of task difficulty with comprehension skill, enjoyment of reading, and knowledge of reading strategies. To account for time on task being nested in students and tasks, random effects for persons and tasks were included. The interaction of task difficulty with gender and Socio-Economic Status (SES) was included for control purposes. Models were estimated individually for 19 countries, and results integrated meta-analytically. In line with predictions, for both time on task indicators, significant positive interactions were found with comprehension skill, enjoyment of reading, and knowledge of reading strategies. These interactions indicated that in students with high comprehension skill, enjoyment of reading, and knowledge of reading strategies there was a stronger association of task difficulty with time on task than in students low in either of these variables. Thus, skilled comprehenders, students enjoying reading, and students in command of reading strategies behaved more adaptively than lower skilled, motivated, or knowledgeable students. Implications of these findings for the validity of self-paced computer-based assessments are discussed.

Keywords: time on task, PISA, educational assessment, test taking motivation, reading skill, reading strategies, validity

INTRODUCTION

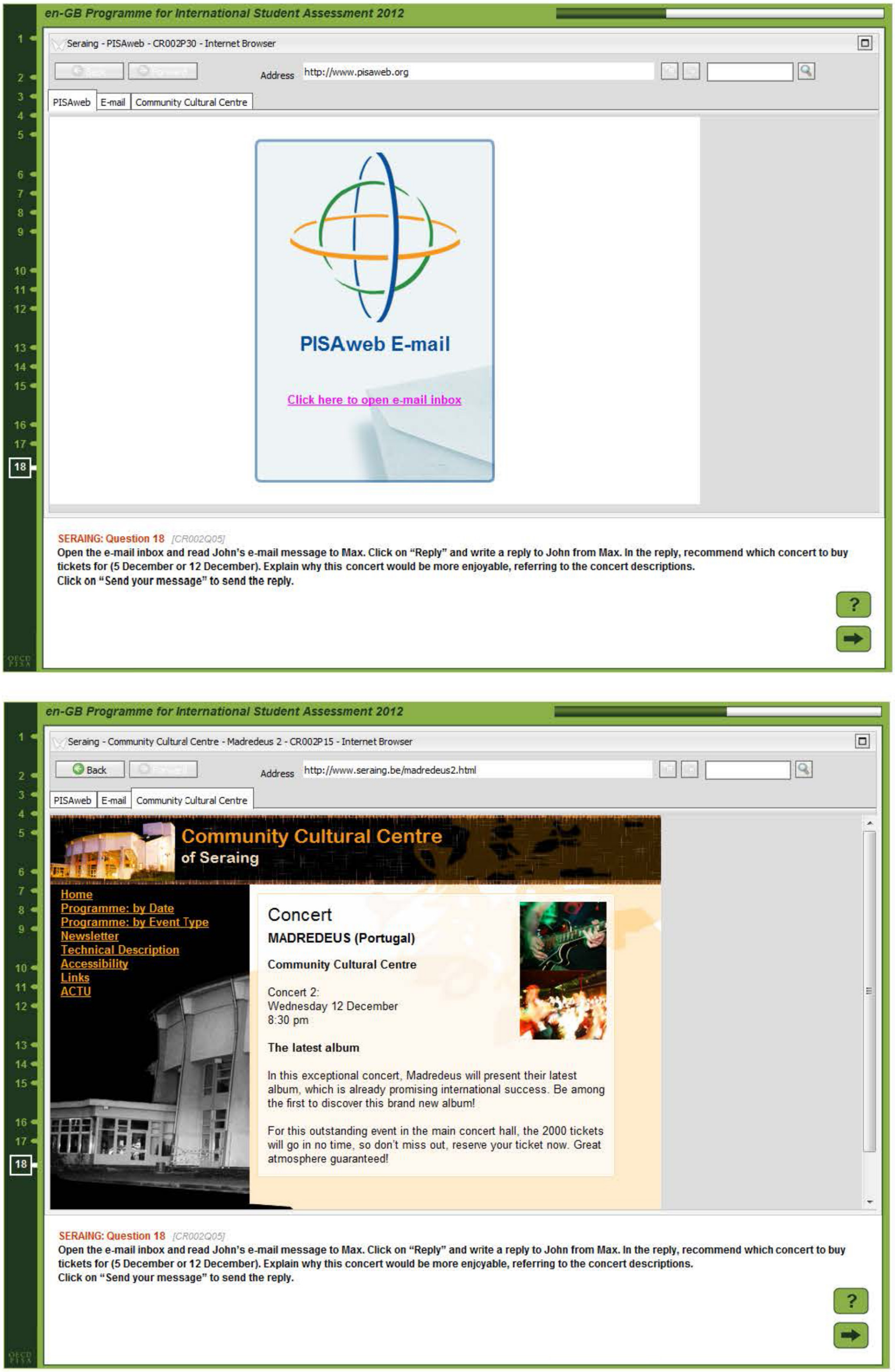
In educational assessments, the goal is to infer a test-taker's latent ability from their performance on a number of tasks. From a psychological perspective however, it is never the latent ability *per se* that determines a test-taker's performance. For the notion of a latent variable to be meaningful, and for the latent variable to be of explanatory value, there has to be some notion of which psychological (and/or neural) mechanisms account for the latent variable taking on a specific value within a specific individual (e.g., Sternberg, 1986; Borsboom et al., 2003). This means that it is always specific cognitive and metacognitive, as well as motivational, processes that are executed during the test takers' engagement with the assessment tasks, which determine the test takers' responses, and thus their estimated abilities. One fundamental process test-takers need to engage in is the allocation of time to individual tasks. This is for two reasons: Firstly, even assessments that are not supposed to be "speeded", i.e., where test-takers are assumed to have ample time to complete all tasks, in fact do have a time limit. Thus, even in these assessments test-takers need to employ some sort of metacognitive strategy to allocate time to individual tasks. Secondly, the time test-takers spend on assessment tasks is a fairly strong predictor of their task performance, where the strength and direction of the association is dependent on characteristics of both the test-taker and the tasks. Apparently, it is especially hard tasks, that cannot be solved by routine cognitive processing, but instead require deliberate, controlled cognitive processing (see Schneider and Shiffrin, 1977; Shiffrin and Schneider, 1977), or metacognitive processing (see Pressley et al., 1989; Winne and Hadwin, 1998) where positive associations between time on task and task performance ("time on task effects") arise. This is e.g., true for tasks from domains such as problem solving in technology-based environments (Goldhammer et al., 2014) or reading digital text (Naumann and Goldhammer, 2017). Against this background, it appears beneficial for a test-taker to invest their time especially in hard tasks. Thus, a natural question seems to be which characteristics of a test taker, either cognitive or motivational, will put them in a position where they adequately allocate their cognitive resources, and thus their time on task, to a task's difficulty. The present research addresses this question for the domain of reading digital text (see e.g., OECD, 2011; Naumann, 2015; Cho et al., 2018). In the following, I will address the ideas that especially students skilled in comprehension ("The skilled"), students knowledgeable of reading strategies ("The knowledgeable"), and students who enjoy reading as such ("The motivated") are successful in adapting the time they invest in a digital reading task to the tasks' difficulty, both overall and regarding the processing of relevant parts of the text materials. These ideas will be derived from describing digital reading as task-oriented reading from the perspective of Rouet et al.'s (2017; see also Britt et al., 2018) RESOLV (REading as problem SOLVing)-model, from Pressley et al.'s (1989) model of the Good Information Processor, as well as the literature on item position effects in assessments (e.g., Debeer et al., 2014), and their moderation through motivation (e.g., Nagy et al., 2018a) and self-control (Lindner et al., 2017).

Comprehension Skill and Task Representation

Reading in an assessment situation is an instance of task-oriented reading (e.g., Vidal-Abarca et al., 2010; Salmerón et al., 2015b; Serrano et al., 2018). In many situations, reading as an activity also is not only the processing of textual information to the end that an adequate situation model of the text contents is being built, as described by cognitive models of text comprehension such as Kintsch's (1998) theory. Rather, especially in opaque information environments such as on line, or when faced with multiple texts that might propose conflicting stances, accomplishing a reading task will entail elements of problem solving (Rouet et al., 2017). When a person reads to solve a task in a reading assessment, they first need to build a representation of the task's requirements. This includes a judgement of whether the question might be answered by a mere memory search (which will not be the case in most reading assessments, which are designed to not rely on prior knowledge). Then, the person will have to judge which parts of the text, or in a multiple text or hypertext reading scenario, which texts are likely to provide the information needed to answer the question. In addition, the task model might include a judgement of the task's difficulty, and thus the required degree of scrutiny in processing the textual information. Consider e.g., the task in **Figure 1**. In this task, students need to compose an e-mail, containing a recommendation to a friend concerning visiting a concert. To accomplish this, students have first to realize that they will need to consult the text. Then, they need to figure out where to find information on the two concerts mentioned in the task instructions, and to match these with the information in the e-mail. As there is no obvious (literal) match between the e-mail and the text on the menu labels in the Seraing Cultural Center's website, they need to figure out a navigation route, finding the Center's program, either by "Date" or by "Event type" to get by the required information. To adequately process this information, they need to figure out they have to evaluate it on a semantic level to judge the concert descriptions against the preferences mentioned in the e-mail. In short, students will have to develop a notion that the task displayed in **Figure 1** is a fairly complex one which requires a good deal of cognitive effort to be solved.

Consider, in contrast, the task displayed in **Figure 2**. Solving this task is possible on the basis of comparatively shallow processing that on a mere lexical level matches the name "Heritage Days" appearing in the question to the same name appearing on the page. The only inferencing needed was due to restrictions on screen resolutions in the assessment, students needed to scroll down to find the relevant information. An appropriate task model in this instance will include the fact that only limited cognitive resources, and time, will be needed to solve it (see also OECD, 2015; Naumann and Goldhammer, 2017).

It is likely that skilled comprehenders will be in a better position to arrive at the judgement that the task displayed in



en-GB Programme for International Student Assessment 2012
Seraing - PISAweb - CR002P30 - Internet Browser
Address
http://www.pisaweb.org
PISAweb
E-mail
Community Cultural Centre
PISAweb E-mail
Click here to open e-mail inbox
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
?

SERAING: Question 18 [CR002Q05]

Open the e-mail inbox and read John's e-mail message to Max. Click on "Reply" and write a reply to John from Max. In the reply, recommend which concert to buy tickets for (5 December or 12 December). Explain why this concert would be more enjoyable, referring to the concert descriptions.

Click on "Send your message" to send the reply.

en-GB Programme for International Student Assessment 2012
Seraing - Community Cultural Centre - Madredeus 2 - CR002P15 - Internet Browser
Back
Address
http://www.seraing.be/madredeus2.html
PISAweb
E-mail
Community Cultural Centre
Community Cultural Centre
of Seraing
Home
Programme: by Date
Programme: by Event Type
Newsletter
Technical Description
Accessibility
Links
ACTU
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
?

Concert

MADREDEUS (Portugal)

Community Cultural Centre

Concert 2:
Wednesday 12 December
8:30 pm

The latest album

In this exceptional concert, Madredeus will present their latest album, which is already promising international success. Be among the first to discover this brand new album!

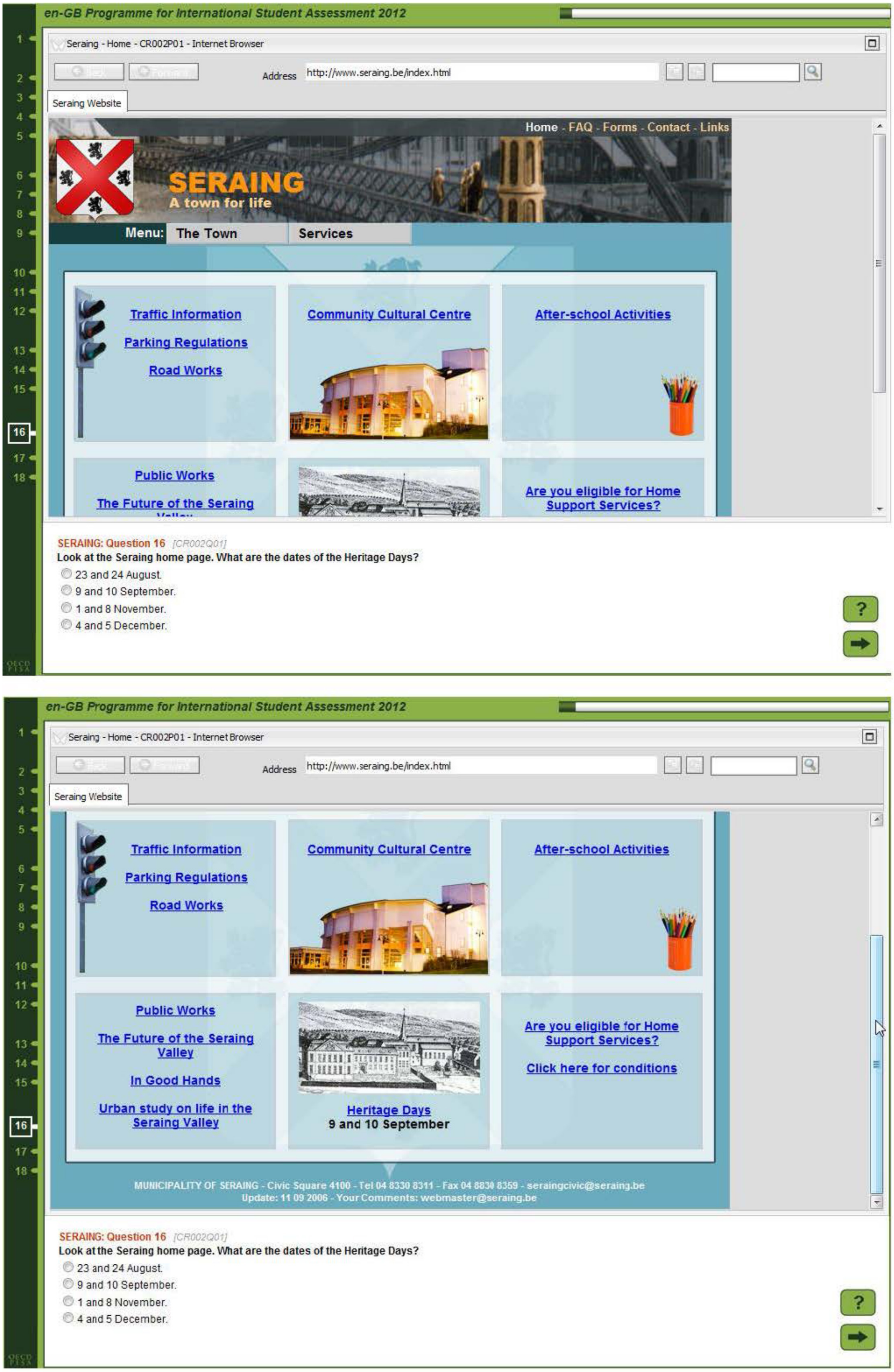
For this outstanding event in the main concert hall, the 2000 tickets will go in no time, so don't miss out, reserve your ticket now. Great atmosphere guaranteed!

SERAING: Question 18 [CR002Q05]

Open the e-mail inbox and read John's e-mail message to Max. Click on "Reply" and write a reply to John from Max. In the reply, recommend which concert to buy tickets for (5 December or 12 December). Explain why this concert would be more enjcyable, referring to the concert descriptions.

Click on "Send your message" to send the reply.

FIGURE 1 | Two screenshots from a digital reading task requiring complex navigation (see OECD, 2011, 2015).


en-GB Programme for International Student Assessment 2012
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
Seraing - Home - CR002P01 - Internet Browser
Back
Forward
Address
http://www.seraing.be/index.html
Seraing Website
Home - FAQ - Forms - Contact - Links
SERAING
A town for life
Menu:
The Town
Services
Traffic Information
Parking Regulations
Road Works
Community Cultural Centre
After-school Activities
Public Works
The Future of the Seraing Valley
In Good Hands
Urban study on life in the Seraing Valley
Heritage Days
9 and 10 September
Are you eligible for Home Support Services?
Click here for conditions
?
OECD PISA

MUNICIPALITY OF SERAING - Civic Square 4100 - Tel 04 8330 8311 - Fax 04 8830 8359 - seraingcivic@seraing.be
Update: 11 09 2006 - Your Comments: webmaster@seraing.be

SERAING: Question 16 *[CR002Q01]*
Look at the Seraing home page. What are the dates of the Heritage Days?

○ 23 and 24 August.
○ 9 and 10 September.
○ 1 and 8 November.
○ 4 and 5 December.

SERAING: Question 16 *[CR002Q01]*
Look at the Seraing home page. What are the dates of the Heritage Days?
○ 23 and 24 August.
○ 9 and 10 September.
○ 1 and 8 November.
○ 4 and 5 December.

FIGURE 2 | Two screenshots from a simple digital reading task requiring minimal navigation (scrolling) only (see OECD, 2011, 2015).

Figure 1 needs ample time to be invested in it, while the task displayed in **Figure 2** might be solved relatively quickly. Similar to the earlier MD-Trace-Model ("Multiple-Document Task-based Relevance Assessment and Content Extraction", see Rouet and Britt, 2011), the RESOLV model postulates a process whereby initially only very coarse reading goals are being set. These reading goals are then constantly updated, and the information acquired is judged against some standard specifying whether enough, and correct, information was acquired to meet the reading goal. According to the standards involved in this process, readers may e.g., judge that they need to re-read a passage, that a passage might be skipped, that it might be sufficient to just skim the passage (e.g., the website in **Figure 2** for the phrase "heritage days"), or that it might be necessary to carefully read a passage, such as the concert descriptions in the task displayed in **Figure 1**.

Previous research has indeed found that skilled comprehenders are better in making decision such as these, compared to lesser skilled comprehenders. One central ingredient of building an adequate task model is to note when, and what, information to search for. In line with the notion that an adequate task model is built more easily by better comprehenders, Mañá et al. (2017) found that decisions to search a text for information was predicted by comprehension skills. Moreover, these authors found that only students with average to good comprehension skills had their search decision, and subsequently task performance, boosted in a condition with a delay between reading the text and reading the questions. In line with these results, Hahnel et al. (2018) found that skilled comprehenders were more likely to seek out additional information when necessary in a task that required the evaluation of on line information provided in Search Engine Results Pages (SERPs).

Again in line with the idea that comprehension skills are a condition for building adequate task models, both Cerdán et al. (2011) and Salmerón et al. (2015a) found that students with higher comprehension skills when studying a text comprising multiple documents were much better in selecting relevant materials, and discarding irrelevant materials. This difference was especially pronounced when there were surface cues present, such as a literal match between a phrase in a passage and in the question, but (other than in the task in **Figure 2**) the passage was in fact irrelevant. Thus, in this scenario, it apparently was good comprehenders who built a task model that (correctly) contained the notion that the surface cue was misleading, and a deeper semantic analysis of the relation between question and text was needed. Similar results were reported by Rouet et al. (2011). These authors found that students in higher grades were less likely to be distracted by semantically irrelevant cues, such as capitalizing, when they had to select hyperlinks from a SERP, than were students in lower grades. A second study showed that indeed parts of this effect could be attributed to students in higher grades having better comprehension skills.

Thus, all in all, if the construction of an adequate task model, that correctly specifies the amount of cognitive effort that has to be invested into a task, is driven by good comprehension skills, we might expect good comprehenders to be better at adapting their time on task to task difficulty in a digital reading situation.

Reading Strategies and Monitoring

As already mentioned in the introductory section of this article, readers in an assessment need to regulate their allocation of time to tasks. Allocating time on task, and monitoring this allocation through the course of completing a reading task can be seen as an instance of the application of cognitive (e.g., planning) and metacognitive (e.g., monitoring) strategies (see Weinstein and Mayer, 1986). As Pressley et al. (1989, p. 858) put it: "Good strategy users employ efficient procedures to accomplish complex, novel tasks... They possess essential metacognitive knowledge for implementing strategies, *including knowing when and where each strategy might be useful, as well as the costs associated with the strategy, such as the amount of cognitive effort it requires*" [emphasis added]. In line with this notion, a number of studies have found that in basic cognitive tasks, subjects tended to align their allocation of time to task difficulty. For instance, Dufresne and Kobasigawa (1989) found that when children in grades 1, 3, 5, and 7 were given a paired association task, where items in one condition were hard (unrelated) and in one condition were easy (related), 5th and 7th graders spent more time on studying the hard, as compared to the easy items, while 1st and 3rd graders showed no such adaptation of study time (see Lockl and Schneider, 2002, for a replication). Consistent with the idea that these differences in study time reflect metacognitive regulation, Lockl and Schneider (2003) demonstrated that indeed judgements of learning ease (estimated effort to learn the items) were higher for hard than for easy items. Consistent with the idea that subjects differ in their ability to effectively regulate their actual study behavior, they also found that 3rd graders showed higher associations between judgements of learning ease, and actual study time than 1st graders.

Such negative associations between judgement of learning ease (the task being perceived as easy) and time on task are however, not uniformly found. For example, Son and Metcalfe (2000, experiment 1), had undergraduate students' study eight biographies of famous people, and answer questions about them. Using these rather complex materials (compared to those used by Dufresne and Kobasigawa, 1989; Lockl and Schneider, 2002), Son and Metcalfe found that students indeed spent *less* time studying the biographies they then judged to be harder. One caveat in this case is however, that judgements of effort were confounded with judgements of interest: Not only were the biographies studied longer that were judged to be easier, but also those that were perceived as more interesting. Thus, it might have been the case that the judgement of effort at least amongst other reflected a lack of interest: The subjectively less interesting biographies were studied quicker, and at the same time judged harder *just because* they were less interesting and thus more effort would have to be put in, to compensate for the lacking interest.

All in all, there appears to be ample, though not unanimous, evidence that students who are able to metacognitively regulate their learning activities spend more time on harder, and less time on easier tasks. There is however, only little direct evidence how knowledge of reading strategies – apart and on top of comprehension skill – would shape the time on task behavior of adolescent students in task-oriented digital reading scenarios, that is, in tasks that are way more complex than even the biographies studied by Son and Metcalfe (2000). Once again from the perspective of the RESOLV-Model, we might expect students knowledgeable of metacognitive reading strategies to be especially apt to align their time on task with task difficulty. This is because during the reading or (in the case of an assessment) task solution process, the task model, i.e., a representation of the reading goal and the resources required and available to achieve it, needs to be constantly updated, and this updating metacognitively regulated (Rouet et al., 2017, see last section, see also Winne and Hadwin, 1998).

Reading Enjoyment and Test-Taking Motivation

Even students who are in good command of comprehension skills, and possess the reading strategy knowledge to successfully build, and through the course of task completion maintain, an adequate task model, might not all alike be motivated to put in the cognitive effort that is required to solve especially hard digital reading tasks. Amongst other lines of research, this is evidenced by studies investigating position effects in low stakes assessments such as PISA. Usually, students' performance declines over the course of an assessment in the sense that the same task will have a lower probability of being answered correctly when it is presented later in the assessment, conditional on a student's skill (Debeer and Janssen, 2013; Debeer et al., 2014; Borgonovi and Biecek, 2016; Weirich et al., 2017; Nagy et al., 2018a). Not all groups of students however are prone to show position effects to equal degrees. For example, Borgonovi and Biecek (2016) analyzed position effects using data from the PISA major domains in 2006, 2009, and 2012, i.e., mathematics, reading, and science, respectively. They found that performance declines due to item positions in each domain to be strong especially in boys, and in students coming from lower socio economic status (SES) backgrounds. Similarly, Nagy et al. (2018b) found that especially for the domain of reading, position effects were strong in boys, and lower SES students.

What mechanisms might account for item position effects in general, and for inter-individual variance in the strength of these effects? The decline in performance in general has been attributed to students, over the course of the assessment, being less willing and/or able to put effort into solving the assessment tasks. For example, Weirich et al. (2017) measured test taking effort at two points in time during 9410 ninth-graders' completion of a science assessment in Germany. They found not only position effects, but these effects, on an individual level, were predicted by the change in test-taking effort that occurred between the two points in time. Lindner and colleagues (Lindner et al., 2017, 2018) discuss position effects in the context of exercising self-control. They define self-control in accordance with Inzlicht and Schmeichel's (2012) process model of self-control. According to this model, exercising self-control at one point in time will decrease especially the motivation to attend to aversive tasks, and increase the likelihood of attendance to pleasing stimuli at a later point in time. Consistent with this idea, Lindner et al. (2018) found that the decline of performance over the course of a 140 min assessment of mathematics and science was predicted by waning state self-control, measured at seven points in time. Also consistent with this idea, Lindner et al. (2017) found that participants who had been forced to exercise self-control in a later mathematics assessment task exhibited a steeper decline in performance (i.e., stronger position effects) than participants who had not had to exercise self-control. However, contrary to their expectations Lindner et al. (2017) did not find any effects of self-control expenditure on time on task as an indicator of task engagement.

According to Inzlicht and Schmeichel (2012), it is especially effort-requiring and for this reason aversive tasks that are affected by previous expenditure of self-control. In the context of cognitive assessments, this assumption implies that waning self-control (and thus a decline in performance) should be strong especially in those students who perceive the assessment tasks as aversive. A reading task, for instance, might be especially aversive for a person who struggles already with basic reading processes, such as decoding, and in general does not enjoy reading. A fluent reader, in contrast, who also enjoys reading as an activity, from this perspective should be much less prone to exhibit position effects. In line with these ideas, Nagy et al. (2018a) indeed found position effects in a reading assessment to decrease with increasing decoding skill and reading enjoyment on the student level.

Taken altogether, we might expect, both from previous research, and from the perspective of theoretical models such as Inzlicht and Schmeichel's (2012) model of self-control, that the adaptation of time on task to task difficulty is dependent not only on cognitive variables such as comprehension skill and knowledge of reading strategies, but also on motivational variables. In particular, we might expect that especially students who perceive reading as an enjoyable activity might be willing to invest extra time when encountering a hard task. Students for whom reading is aversive, in contrast, might refrain from this investment, so that the adaptation of time on task to task difficulty should be especially pronounced in motivated readers, who report a high level of reading enjoyment.

The Present Research

To the best of the author's knowledge, there is yet no study that investigates how in reading digital text, students' adaptation of total time on task to task difficulty is conjointly predicted by comprehension skill, knowledge of reading strategies and reading enjoyment. In task-oriented reading of multiple texts in general, and in task-oriented reading situations using digital text in particular, readers need to select which texts, or parts of the text available, to access and to use, in which order to accomplish their goals, and which to discard ("navigation", see Lawless and Schrader, 2008; Naumann, 2015; Salmerón et al., 2018). Then they have to decide for each text or part of a text selected, how much cognitive effort they want to invest into processing. Naturally, especially in hard tasks, it seems beneficial to devote time to processing task-relevant parts of the available materials (see e.g., Rouet and Le Bigot, 2007). Thus, besides investigating the differential adaptation of total time on task to task difficulty, the present research specifically examined how the time students spend on relevant parts of the text stimulus is adapted to task difficulty by students varying in comprehension skill, knowledge of reading strategies, and reading enjoyment. These questions are addressed using data from one of the first computer-based large-scale assessments, the PISA 2009 Digital Reading Assessment.

The Digital Reading Assessment was an International Option in PISA 2009, which was chosen by 19 countries and economies. It was targeted specifically at students' skill in engaging with, comprehending, and using digital texts that were prevalent at the time the assessment was conceived (early 2007 to early 2008), such as websites (personal, educational or corporate), blogs, e-mails, or forums. It was comprised of a total of 29 tasks, which were distributed across nine units. Each unit consisted of a text stimulus and between one and four tasks. Each text stimulus was made up of several pages, which in most cases belonged to different texts, such as an e-mail and a website (see **Figure 1**). Tasks differed in how many pages students needed to access to complete the task, with some tasks requiring to read only the task's prompting page (see **Figure 2**), and some tasks requiring the student to perform as many as 13 steps of navigation. Besides pages necessary to complete the task, tasks also varied in their number of relevant pages. Relevant pages were defined as those pages that either contained information that needed, or could be used to solve each task, or that needed to be visited in order to arrive at this information. In addition, pages were considered relevant that, from their labels, could be assumed to hold information instrumental either to solve the task, or to complete navigation, such as a "site map". The mean number of relevant pages was 3.61 (*SD* = 3.42, *Md* = 2, *Min* = 1, *Max* = 14). However, in each task, all pages of the unit's text stimulus were available to students, making it possible to visit not only pages that were relevant to the task, but also non-relevant pages.

The PISA 2009 Digital Reading data set lends itself to address the issues raised in a couple of ways. First, computer-based assessments allow for the measurement of time on task, and more so, for a detailed investigation of what parts of a task stimulus (in this case: the text[s]) students encountered for how long, and in which sequence. This makes it possible to derive measures of task engagement, such as the average time spent on relevant pages, which are not routinely available from paper and pencil tests (see Greiff et al., 2015). Second, a number of tasks large enough to model a random effect for tasks is available. Thus, other than in fixed effects models such as ANOVA or OLS regression, which allow generalization only to other persons, but not to other situations, conditions, or tasks than those specifically employed in the respective design, here the obtained results can in principle be generalized to other tasks that were constructed according to the same framework through modeling task as a random effect (De Boeck, 2008). Third, since reading was a major domain in PISA 2009, rather detailed student-level measures are available, not only as to their comprehension skill, but also as to their knowledge of reading strategies, and their enjoyment of reading. Finally, large scale databases provide not only good variation in terms of students' backgrounds, but also good opportunities to control for background variables such as SES and gender. In the present case, this seems especially crucial, as on the grounds of the results reported by Nagy et al. (2018b) and Borgonovi and Biecek (2016, see section "Reading Enjoyment and Test-Taking Motivation" above), it might well be expected that higher SES students and girls are more likely to adapt their time on task behavior to task difficulty than are their lower SES peers or boys: As it seems, higher SES students, as well as girls, are more prepared than their lower SES or male peers to maintain cognitive effort in an assessment. This means that these background variables also are likely to affect students' preparedness to adapt their time on task behavior to task difficulty. Thus, any analysis targeting time on task behavior conditional on task difficulty should control for the interaction of SES and gender with task difficulty.

MATERIALS AND METHODS

Subjects

Subjects were those students that participated in the PISA 2009 Digital Reading Assessment and for whom time on task for at least two tasks, comprehension skill, knowledge of reading strategies, and enjoyment of reading were available ($N = 32{,}669$, country-wise $930 \leq N \leq 2800$, see **Supplementary Material 1** for country-wise *N*'s). Overall, there were 50% boys. There were between 46 and 53% boys in each sample. Due to PISA's sampling scheme, which samples students at the end of compulsory education, students were between 15.17 and 16.33 years old (*M* = 15.78, *SD* = 0.29; country-wise *M* between 15.67 and 15.87).

Measures

Total Time on Task

Time on task was read from log files. It was defined as the time that elapsed between the onset of the task, and the time the student gave a response. It thus comprised the time a student spent reading the task instruction, reading potentially both relevant and irrelevant parts of the text, and deciding on a response. To account for the skew of the time on task distribution, the natural logarithm of the total time on task was used.

Time on Relevant Pages

To compute time on relevant pages, each navigation sequence was segmented by page transitions. Then the time elapsed between each transition to, and from, a page classified as task-relevant was summed up across each task-completion sequence. Since in each task the prompting page was defined as relevant, time on relevant pages also comprised the time spent reading the task instruction. It did however not comprise time a student might have spent reading task-irrelevant parts of the stimulus. Because tasks varied considerably in the number of relevant pages they comprised, time on relevant pages was standardized at the number of relevant pages available in each task. To account for the skewness of the distribution, the natural logarithm of time on relevant pages was used.

Comprehension Skill

Comprehension skill was measured through the PISA 2009 print reading assessment. Being a major domain in 2009, print reading skill was measured with a total of 131 items in 37 units (a unit consists of a text stimulus accompanied by either a single, or multiple items). These 131 items were allocated to 13 clusters worth of approximately 30 min of testing time each. The clusters were assigned to 13 different booklets together with items from the PISA mathematics and science assessments. Each booklet contained four clusters. Of these 13 booklets, one contained four clusters of reading items, three contained three clusters of reading items, seven contained two clusters of reading items, and two contained one cluster of reading items. Thus, each student completed at least 30 min of print reading, with 12 out of 13 students completing at least 60 min (see OECD, 2012, p. 29–30 for details). Items had been constructed according to an assessment framework (OECD, 2009) specifying three different reading aspects, or cognitive operations: (1) Accessing and retrieving, (2) integrating and interpreting, and (3) reflecting and evaluating textual information, as well as two different text formats, continuous and non-continuous texts (see OECD, 2009). It is important to note that in both continuous and non-continuous texts in the print reading assessment students were prompted with the complete text, thus, no navigation in the sense of physical access to text through hyperlinks was required. Comprehension skill was scaled according to the Rasch Model. Weighted Maximum Likelihood Estimates (WLEs) were used in the present analysis. The WLE reliability was 0.84 (see OECD, 2012, p. 194, Table 12.3).

Knowledge of Reading Strategies

Knowledge of reading strategies was measured with two reading scenarios. In each scenario students were prompted with a specific reading situation. These reading situations were the following: (a) "You have just read a long and rather difficult two-page text about fluctuations in the water level of a lake in Africa. You have to write a summary", and (b) "You have to understand and remember the information in a text". Each of these reading scenarios were accompanied by either 5 (summary scenario) or 6 (understanding and remembering scenario) possible strategies such as "I try to copy out accurately as many sentences as possible" (summary) or "I quickly read through the text twice" (understanding and remembering). In each scenario, each strategy had to be rated by students on a 6-point rating scale from "not useful at all" to "very useful". It is important to note that the students did not rank-order the strategies themselves, but rated them for their usefulness independently from each other", and these ratings were then in a second step ranked-ordered within each scenario and student. At the same time, the strategies had been rated, and rank-ordered, by reading experts. The scoring then was accomplished on the basis of the agreement between the rank-order of each student's ratings with the experts' ratings' rank-order. Specifically, 1 point was awarded for each pairwise comparison in students' ratings that agreed with the respective pairwise comparison in the experts' rating for those 9 (understanding and remembering) and 8 (summarizing) pairs of strategies where there was consensus amongst the experts which strategy was more useful. A point was only awarded when students, in agreement with experts, ranked a strategy to be *more* useful than another. Thus, when two strategies that entered the score were ranked as *equally* useful by a student, no point was awarded (see OECD, 2012, p. 282). The possible score thus ranged between 0 (no agreement) and 17 (agreement in all 17 pairwise comparisons considered).[1] The reliability (Cronbach's α) for the 17 pairwise comparisons entering the score was 0.84 in the present sample, the EAP reliability was 0.86.

Reading Enjoyment

Enjoyment of reading was measured through 11 items such as "Reading is one of my favorite hobbies" or "For me, reading is a waste of time", which were to be answered on a 4-point Likert scale ranging from "Strongly disagree" to "Strongly agree". Item wordings and item parameters can be found in OECD (2012, p. 290). For the present research, the enjoyment of reading index provided in the OECD PISA 2009 data base was used. Reading enjoyment was scaled according to the partial credit model, providing a Weighted Maximum Likelihood Estimate (WLE) for each student. The reliability (Cronbach's α) for the present sample was 0.89.

Task Difficulty

Task difficulty was defined using the item difficulties of the PISA 2009 digital reading items. In PISA, items are scaled according to the Rasch model. The simple logistic model is applied to dichotomous items, while partial credit items are scaled according to the partial credit model (Masters, 1982). Of the 29 reading tasks in the digital reading assessment, eight had partial credit. Item difficulties (delta) were taken from the international calibration of the PISA 2009 digital reading items, which are provided in OECD (2012, Table A4, p. 343). For partial credit items, this parameter marks the location of the latent ability continuum where the likelihoods of a responses in the highest and the lowest response category are equal (see e.g., Adams et al., 2012).

Socio-Economic Status (SES)

To measure students' SES, the PISA ESCS index was used, which is composed of students' parents' occupational status, students' parents' education, and wealth, as well as cultural and educational resources in students' homes (including, but not limited to, the number of books at home). Technically, the ESCS is a factor score from a principal component's analysis of the HISEI (highest parental occupation amongst a student's parents), and the PISA home possessions index (HOMEPOS). Details on how the ESCS was computed in PISA 2009 can be found in OECD (2012, p. 312–313).

Procedure

Students were tested in schools during school hours. First, students completed the paper-based cognitive assessment

[1] Note that in PISA 2009, two separate indices were built on the basis of the two scenarios. In the present research, the two scenarios were combined into one score in accordance with the intentions of the authors of the original instrument from where idea of measuring strategy knowledge employed in PISA 2009, as well as the scenarios and to-be-rated strategies were derived (see Artelt et al., 2009).

(reading, mathematics and science), which lasted for two hours. Students could take a break after one hour. Afterwards, the student questionnaire was administered. Last, students completed the computer-based reading assessment. In PISA 2009 digital reading skill was the only domain in the computer-based assessment. Digital reading items were presented in a secure test environment where a browser was simulated that had all typical features of commercial web browsers at the time the assessment was conceived. Items were presented unit by unit, and in each item, the unit's text(s) were accessible, regardless of whether they were relevant to the item at hand or not. After giving a response, students could not go back to correct their response. Testing time in the Digital Reading Assessment was 40 min. Students knew in advance how much time in total there was to complete the assessment. In addition, students first completed a 10-min tutorial where they could make themselves familiar with the testing environment and simulated web browser. The assessment was not speeded, as indicated by a small number of not-reached items (0.4 on average, see OECD, 2012, chapter 12).

All testing and other data collection instruments and procedures were approved by the PISA governing board, composed of country representatives of all countries that participated in the assessment, as well as by the PISA consortium, led by the Australian Council for Educational Research. Implementation of data collection and management was overseen by national centers, led by national project managers, in each country (see OECD, 2012, p. 24–25 for details). The data that are used for the present research are either in the public domain, and can be found at http://www.oecd.org/pisa/data/ (accessed March 01, 2019), or, where this was not the case, the author had received written consent from OECD to use the Digital Reading Assessment log file data for scientific purposes to be published in scholarly journals. An ethics approval was thus not required for this study as it presents a secondary analysis of OECD data. The author of the present article at no point had access to information identifying individual subjects.

Statistical Modeling Approach

Linear Mixed Model and Estimation

To account for item-specific response times being nested both in items and students, a linear mixed model (LMM) framework was employed that specified crossed random effects for student and item intercepts, and an additional random effect for schools to account for the fact of students being nested in schools due to the PISA sampling procedure. The central research questions were addressed by regressing time on task on the student level variables comprehension skill, knowledge of reading strategies, and reading enjoyment, and the task-level variable task difficulty, as well as, most importantly, the interaction of each student level variable with task difficulty. On top of the main effects and the three two-way interactions of comprehension skill, knowledge of reading strategies, and reading enjoyment with task difficulty, the model contained all other possible two, three and four-way interactions between the four theoretically relevant variables. Gender and SES were entered as control variables. Since the theoretically relevant effects were two-way interactions involving task difficulty, the two-way interaction of each gender and SES with task difficulty was entered into the model as well. No other or higher-order interaction terms involving gender and SES were specified.

All models were estimated in the R environment (R Development Core Team, 2008) using the function lmer from the package lme4 (Bates et al., 2015), version 1.1-15. For better interpretability of regression coefficients, all metric variables were centered and standardized within each country or economy. This means that regression coefficients represent expected changes in the criterion variable in terms of its within-country standard deviation, per within-country standard deviation of each predictor. Standard deviations of all variables in the analyses did not vary much across countries (see **Supplementary Material 1**). Gender was entered dummy-coded with girls as the reference group.

Integration of Country-Specific Results

Country-specific results (fixed effects) were integrated using a random-effects meta-analytic model (Hedges and Vevea, 1998), using the R-package metafor (Viechtbauer, 2010). Meta-analysis lends itself for the analysis of data such as the present for multiple reasons. In educational assessments such as PISA, sampling occurs at the level of countries, so that an analysis pooling data from all countries would not be appropriate. However, besides effects for individual countries, it is of interest how an effect turns out in general, i.e., across countries. A random-effects meta-analytic model that discriminates a fixed (total) effect from a random, study-specific effect seems especially suitable in this situation: The fixed effect may be interpreted as a general effect, which is the same across countries. The variance of the study (i.e., country) specific effect gives an estimate, and allows a significance test, for the variance of county specifics adding to the total effect size, over and above sampling variance.[2] To conduct the meta-analysis for each effect, one vector was created for each effect containing the country-specific estimates of each effect through reading the respective effect from the respective lmer object using the function fixef from the lme4 package. A second vector containing each effect's standard error for each country was created using the se.fixef function from the package arm (Gelman and Su, 2016). These two vectors (after taking the square of each effect's standard error to arrive at the variance) were given to the rma function from the metafor package. An alpha level of 0.05 was set for all significance tests.

[2] Theoretically, an alternative to the meta-analytic approach used here would have been a model where country is treated as a random effect, and a random slope is estimated for each effect across countries. However, apart from the fact that given the number of fixed effects in the present analysis, such a model would probably would have been computationally intractable, it would only tell us *if* an effect varies across countries, but not in which way. Including country as another fixed effect, and estimating its interaction with each of the other fixed effects in the analysis would have added at least another 20 fixed effects to an already complex model. Thus, in the present case, the meta-analytic approach appeared to be the best compromise between comprehensiveness and parsimony that could be found.

Illustration of Interaction Effects Through Simple Slopes

To illustrate the interaction effects between task difficulty and comprehension skill, knowledge of reading strategies, and reading enjoyment, respectively, simple slopes were computed and tested for significance at the upper and lower boundaries of the respective distributions (2.5th and 97.5th percentiles, or ±1.96 standard deviations). For comprehension skill (and task difficulty) these percentiles represent the boundaries between the highest and second to highest competency level (levels 5 and 6), and the lowest and second to lowest competency level (levels 1a and 1b) respectively (see OECD, 2010, for the interpretation and description of reading competency levels). The values at which to compute simple slopes were chosen for knowledge of reading strategies and reading enjoyment in accordance. It is important to note that irrespective of the values chosen for the computation of simple slopes, the interaction effect as such relates to the whole sample, and simple slopes could, in principle be computed for *any* value of each predictor in the model (see Aiken et al., 2003).

RESULTS

Means, standard deviations and correlations for all variables in the analyses pooled across countries and economies are provided in **Table 1**. Country-specific statistics are provided in **Supplementary Material 1**.

Random Effects

There was significant variation of time on task, as well as time on relevant pages, between tasks, subjects, and schools in each country and economy. The corresponding variance components can be seen in detail in the model summaries that are provided as **Supplementary Material 2**. **Supplementary Material 3** provides the respective significance tests. In the following, all estimates are meta-analytic fixed effects across countries and economies. Country-specific effects can be found in **Supplementary Material 2**. Most of the fixed effects of theoretical interest showed significant variability across countries and economies, over and above sampling variance. Since, however, this variability in the present research was not of theoretical interest, the estimates and the significance of between-country variance is presented as **Supplementary Material 4**. In the following, if a fixed effect showed *no* variance across countries over and above sampling variance (the exception from the rule), this is explicitly mentioned.

Fixed Effects

Main Effects of Task Difficulty Comprehension Skill, Strategy Knowledge and Reading Enjoyment

As expected, there was a significant main effect of task difficulty, meaning that students on average took more time in harder tasks (meta-analytic effect: b = 0.39, SE = 0.02, 95%-CI: [0.35; 0.43]), and on average spent more time on task-relevant pages (meta-analytic effect: b = 0.18, SE = 0.03, 95%-CI: [0.13; 0.24]). Neither main effect of task difficulty varied across countries over and above sampling variance. Also, both time on task indicators were positively predicted by comprehension skill. More skilled comprehenders spent more time on the tasks in general, and they spent more time on relevant pages (meta-analytic effect for both time on task indicators: b = 0.09, SE = 0.01, 95%-CI: [0.08; 0.10]).

On top of the main effect for comprehension skill, there was a positive main effect of strategy knowledge on both time on task indicators. For both time on task indicators this effect was b = 0.03 (SE < 0.01), 95%-CI: [0.02; 0.03]. In addition to the main effects of comprehension skill and strategy knowledge, reading enjoyment had a positive main effect, meaning that students enjoying reading both spent more time on the tasks in total (meta-analytic effect: b = 0.02, SE < 0.01, 95%-CI: [0.01; 0.03]), and on relevant pages (meta-analytic effect: b = 0.02, SE < 0.01, 95%-CI: [0.01; 0.02]).

Interactions of Task Difficulty With Comprehension Skill, Strategy Knowledge and Reading Enjoyment

Comprehension Skill

The main effects of task difficulty and comprehension skill were qualified by a significant positive two-way interaction (see **Figure 3**, left panel, and **Figure 4** for an illustration). Meta-analytically, this interaction amounted to b = 0.09 (SE < 0.01), 95%-CI: [0.08; 0.09] for total time on task, and b = 0.08 (SE < 0.01), 95%-CI: [0.07; 0.08] for time on relevant pages (see the left hand panel in **Figure 3**), representing a medium-sized effect each.

TABLE 1 | Means, standard deviations, and correlations for all variables in the study in their original metric and coding.

	Min	Max	*M*	SD	Correlations						
(1) Total time on task[a,b]	1.17	1753.42	104.21	86.23	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(2) Time on relevant pages[a,b,c]	0.07	1753.42	42.57	46.59	0.42						
(3) Task difficulty[a]	−2.72	2.33	−0.01	1.04	0.38	0.21					
(4) Comprehension skill[d]	0.00	884.66	501.28	100.70	0.08	0.03	0.00				
(5) Strategy knowledge[d]	0.00	17.00	9.86	4.39	0.06	0.03	0.00	0.43			
(6) Reading enjoyment[d]	−3.23	3.49	0.02	0.98	0.06	0.04	0.00	0.39	0.27		
(7) Gender[d,e]	1.00	2.00	1.49	0.50	−0.04	−0.03	0.00	−0.17	−0.16	−0.28	
(8) SES[d]	−6.04	3.03	−0.06	0.99	−0.01	−0.03	0.00	0.33	0.22	0.13	0.01

[a] *k = 640,482 task responses.* [b] *Seconds.* [c] *Averaged across the number of relevant pages available.* [d] *n = 32,699 students.* [e] *1 = female, 2 = male.*

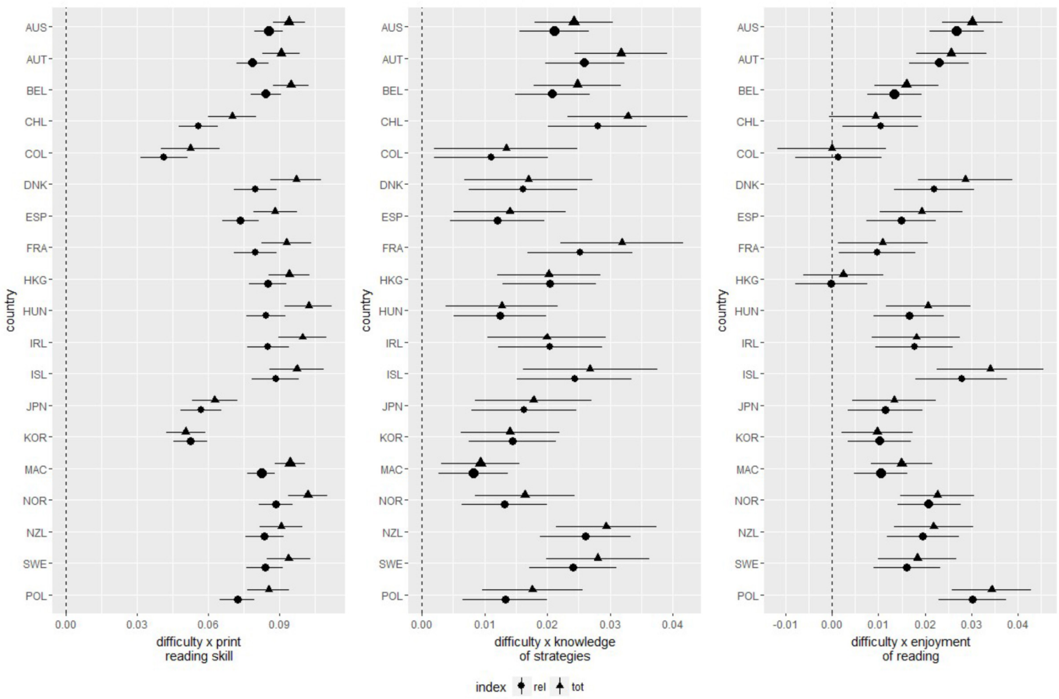


FIGURE 3 | Interaction of task difficulty with comprehension skill, knowledge of reading strategies, and reading motivation as predictor of total time on task (tot) and average time on relevant hypertext pages (rel). Error bars indicate 95% confidence intervals. Symbol sizes are proportional to precision of each estimate.

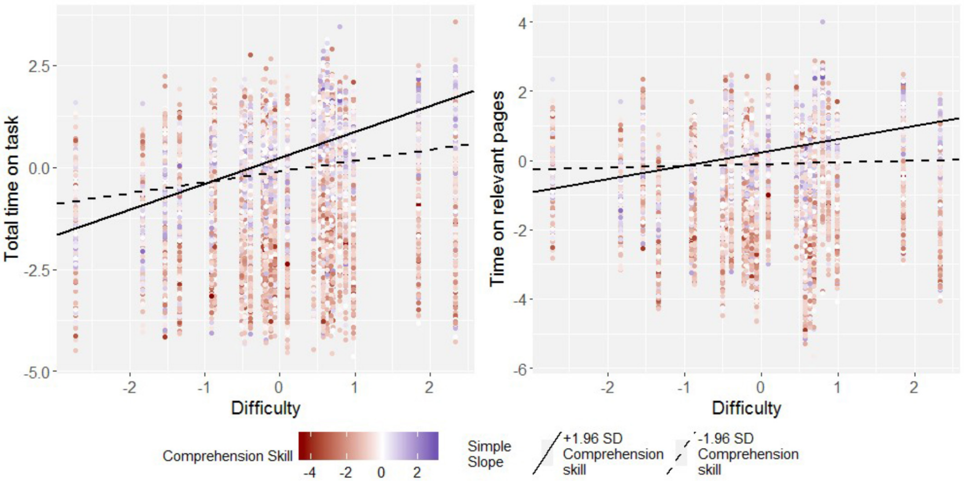


FIGURE 4 | Simple slopes for the regression of total time on task and time on relevant pages on task difficulty in students high (97.5th perc.) and low (2.5th perc.) in comprehension skill for one sample country (Australia). Data points are raw data. Regression intercepts and slopes are model-based estimates.

To further interpret these interactions, simple slopes were computed depicting the effect of task difficulty in very strong comprehenders and very weak comprehenders ($z_{\text{comprehension}} = \pm 1.96$, see **Figure 4** for an illustration). Likewise, the effect of comprehension skills was estimated in very hard and very easy items ($z_{\text{difficulty}} = \pm 1.96$). These analyses revealed the following: There was a strong effect of task difficulty on both time on task indicators in strong readers, amounting meta-analytically to $b = 0.60$ ($SE = 0.02$), 95%-CI: [0.52; 0.61] for total time on task, and to $b = 0.33$ ($SE = 0.03$), 95%-CI: [0.28; 0.39] for time on relevant pages. Both these slopes had no significant variance across countries. For poor comprehenders at the lower end of the comprehension skill distribution the effect of task difficulty on total time on task was much reduced, though still significant, the meta-analytical effect was $b = 0.22$ ($SE = 0.02$), 95%-CI: [0.18; 0.26]. No significant effect of task difficulty on time on relevant pages was found in poor comprehenders, b = 0.03 (SE = 0.03), 95%-CI: [−0.02; 0.09]. Once again, these two simple slopes displayed no variance over and above sampling variance.

Correspondingly, in hard tasks, there was a strong positive association of comprehension skill with both total time on task, b = 0.26 (SE = 0.01), 95%-CI: [0.23; 0.28], and time on relevant pages, b = 0.23 (SE = 0.01), 95%-CI: [0.22; 0.26]. In easy tasks, in contrast, this association was negative for both total time on task, b = −0.08 (SE = 0.01), 95%-CI: [−0.10; −0.07], and time on relevant pages, b = −0.06 (SE = 0.01), 95%-CI: [−0.07; −0.05].

Taken altogether, these results suggest the following: Skilled comprehenders align their total time on task, as well as the time they spend on task-relevant hypertext pages, closely to the tasks' difficulties. In contrast, much less of such an adaptive behavior occurs in poor comprehenders. These readers show some alignment of their total time on task with task difficulty, but none of the time they spend on relevant parts of the text. Correspondingly, when tasks were hard, skilled comprehenders appeared to invest more time in these tasks than poor comprehenders. Easy tasks in contrast were more quickly solved by skilled, as opposed to poor comprehenders.

Knowledge of Reading Strategies

The positive main effect of strategy knowledge on both total time on task and time on relevant pages was in each case qualified by a significant positive interaction with task difficulty, amounting to b = 0.02 (SE < 0.01), 95%-CI: [0.02; 0.02] both for total time on task and time on relevant pages (see the middle panel in **Figure 3**, and **Figure 5** for an illustration), which represented a small effect each. To interpret theses interactions, simple slopes were computed to estimate the effect of task difficulty for students at the upper and lower ends of the strategy knowledge distribution ($z_{\text{strategyknowledge}}$ = ± 1.96), and, correspondingly, the effect of strategy knowledge in easy and hard items. For students high in knowledge of reading strategies, the effect of task difficulty on total time on task was estimated as b = 0.44 (SE = 0.02), 95%-CI: [0.40; 0.47], and on time on relevant pages as b = 0.22 (SE = 0.03), 95%-CI: [0.16; 0.29]. Both these effects were homogeneous across countries and economies. For students low in knowledge of reading strategies, the effects of task difficulty on time on task were still significant, but reduced in magnitude. They amounted to b = 0.35 (SE = 0.02), 95%-CI: [0.31; 0.39] for total time on task and b = 0.15 (SE = 0.03), 95%-CI: [0.09; 0.21] for time on relevant pages. Once again, these two effects showed no variability over and above sampling variance across countries and economies.

In hard tasks, the effect of strategy knowledge on total time on task was estimated as b = 0.07 (SE = 0.01), 95%-CI: [0.06; 0.08], and the effect on time on relevant pages as b = 0.06 (SE < 0.01), 95%-CI: [0.05; 0.07]. These positive associations were reversed to negative in easy tasks, where the effect of strategy knowledge on total time on task was b = −0.02 (SE < 0.01), 95%-CI: [−0.01; −0.02], and on time on relevant pages b = −0.01 (SE < 0.01), 95%-CI: [−0.01; 0.00].

Taken together these results suggest that over and above the effect of comprehension skill, students with better knowledge of reading strategies do a better job in aligning their time on task behavior with task difficulty. Students with better knowledge of reading strategies at the same time invest more time in hard tasks, and are quicker in solving easy tasks, than their less knowledgeable peers.

Reading Enjoyment

As were the main effects of comprehension skill and knowledge of reading strategies, the main effect of reading enjoyment was moderated by task difficulty though a significant positive interaction, b = 0.02 (SE < 0.01), 95%-CI: [0.01; 0.02] for both total time on task and time on relevant pages (see the right hand panel in **Figure 3**, and **Figure 6** for an illustration), which represented a small effect each. Simple slopes analyses (see **Figure 6** for an illustration) revealed that in students high

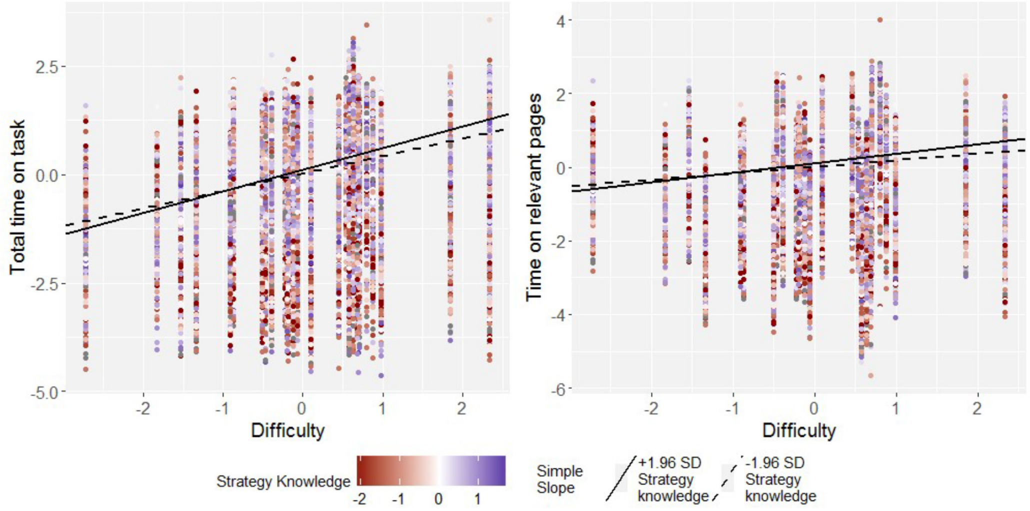


FIGURE 5 | Simple slopes for the regression of total time on task and time on relevant pages on task difficulty in students high (97.5th perc.) and low (2.5th perc.) in knowledge of reading strategies for one sample country (Australia). Data points are raw data. Regression intercepts and slopes are model-based estimates.

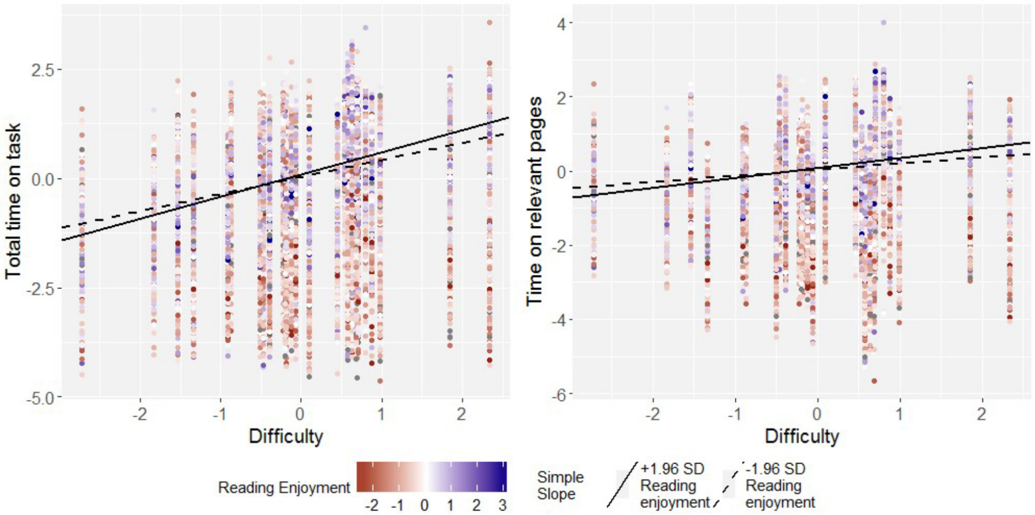


FIGURE 6 | Simple slopes for the regression of total time on task and time on relevant pages on task difficulty in students high (97.5th perc.) and low (2.5th perc.) in enjoyment of reading for one sample country (Australia). Data points are raw data. Regression intercepts and slopes are model-based estimates.

in reading enjoyment, there were strong or medium sized effects of task difficulty on both total time on task, $b = 0.43$ ($SE = 0.02$), 95%-CI: [0.39; 0.47], and time on relevant pages, $b = 0.22$ ($SE = 0.03$), 95%-CI: [0.16; 0.27]. These effects were reduced, but remained positive and significant in students low in reading enjoyment, where they amounted to $b = 0.36$ ($SE = 0.02$), 95%-CI: [0.32; 0.40] for total time on task, and $b = 0.16$ ($SE = 0.03$), 95%-CI: [0.09; 0.21]. All simple slopes for task difficulty in students low and high in reading enjoyment did not display variance over and above sampling variance.

As for comprehension skill and knowledge of reading strategies, a positive effect of enjoyment of reading was found in hard tasks, which amounted to $b = 0.05$ (SE = 0.01), 95%-CI: [0.04; 0.06] for both total time on task and time on relevant pages. In easy tasks, this effect was once again reversed to negative, and amounted to $b = -0.02$ ($SE < 0.01$), 95%-CI: [−0.03; 0.00] for total time on task, and $b = -0.01$ ($SE < 0.01$), 95%-CI: [−0.02; 0.00]. Thus, on top of the corresponding effects for comprehension skill and knowledge of reading strategies, students who enjoy reading more appear to invest more time in difficult tasks, but are quicker when they work on easy tasks than their peers who report less enjoyment in reading. It should be noted though that the negative effect of enjoyment was small, and that simple slopes were computed for tasks at the lower and upper end of the task difficulty distribution. Thus, for easy to moderately difficult tasks, the effect for enjoyment in reading on time on task would be zero, or slightly positive.

DISCUSSION

The present article examined the task-adaptive allocation of time, and time spent on relevant pages, while reading digital text, dependent on students' comprehension skills, knowledge of reading strategies, and enjoyment of reading. Although these three student characteristics are positively correlated (see **Table 1** and **Supplementary Material 1**), independent effects (that is, while controlling for each other) could be secured, indicating that students high in each of these variables showed a more pronounced adaptation, both of total time on task and of time on relevant pages, to the tasks' difficulties. This was evidenced by significant positive interactions of each these student characteristics with task difficulty in predicting time on task and time on relevant pages, which were found consistently across 19 countries and economies (the only exception being Colombia and Hungary, where no interaction of task difficulty with reading enjoyment was found, see **Figure 3**).

The Present Results Viewed From Previous Theory and Findings

These results are much in line with research from cognitive, educational, and social psychology that describes how students build models of the task when reading, how they monitor the reading process, and how they maintain effort when encountering a lengthy assessment comprising of multiple tasks, such as PISA. Specifically, the finding that time on task and task relevant pages are more positively predicted by task difficulty in strong comprehenders is much in line with the RESOLV model (Rouet et al., 2017), as strong comprehenders can be expected to be better in creating adequate task models. It is also in line with previous research pointing to better comprehenders behaving more task-adequate when it comes to selecting relevant, and discarding non-relevant text materials (Cerdán et al., 2011; Salmerón et al., 2015a). The result that knowledge of reading strategies is predictive of the adaptivity of time on task behavior also is in line with the RESLOV model, as well with earlier models of metacognitive engagement while learning, such as Winne and Hadwin's (1998) COPES (Conditions, Operations, Procedures, Evaluations, Standards) model. Finally, the interaction of task difficulty with reading enjoyment is consistent with research describing position effects, or performance declines, in low-stakes assessments as a result of failing self-control and, as a result, motivation to mobilize mental resources (e.g., Lindner et al., 2017; Nagy et al., 2018a). These effects are moderated by students' enjoyment of reading, presumably because these students view the assessment task as less aversive, and thus suffer less from failing self-control (Inzlicht and Schmeichel, 2012). From this perspective, it was to be expected, that students enjoying reading as an activity would also be more likely to invest time especially in hard tasks. This latter result is also nicely aligned with recent descriptions of "engaged" reading as proposed by Guthrie and colleagues (Guthrie et al., 2012). In their model, a direct predictor of reading achievement is behavioral engagement, which they also coin "dedication" (p. 604). Behavioral engagement in itself is dependent on motivations to read. In the present context, we might well assume behaviorally dedicated students especially those who allocate their time especially in hard tasks, and devote extra time especially to reading relevant parts of the text when the task is hard.

Implications of the Present Results for Assessment and Education

As mentioned in the introductory part of this article, completion of an assessment task, and, in turn, the estimated ability of a student is not merely the reflection of a latent variable. Rather, it is always the result of intertwined cognitive and motivational processes carried out at time of task completion. One of these processes is the task-adequate mobilization of cognitive resources, and thus the expenditure of time. From the perspective of the present results thus the question arises whether time on task, or time spent on relevant pages, is governed by variables that can be regarded as part of the to-be-measured construct "digital reading skill". In other words: If a crucial process of task engagement, that is predictive of task performance, is functionally dependent on processes and dispositions that are clearly outside the definition of the targeted construct, this would pose a threat to validity arguments made on the basis of the respective test scores (AERA et al., 2014). The largest interaction effects found in the present research were those of task difficulty with comprehension skill. Comprehension however clearly is part of the construct "digital reading", as digital reading is reading in the first place. Thus, if a student is in a better position to solve a digital reading task due to better comprehension skill in part because these superior comprehension skills enable them to better align their effort with the task's requirements, this does not necessarily pose a threat to the assessment's validity. Rather, one might argue, it describes an additional pathway whereby good comprehension skills predict good performance in digital reading, and thus explain the positive correlation that is usually found between offline and online measures of reading skill and performance (e.g., Coiro, 2011; OECD, 2011; Naumann and Salmerón, 2016).

A similar argument might be made for knowledge of reading strategies. A long tradition of previous research has pointed to the necessity of strategic control especially in reading situations encountering digital text, web-based text, hypertext, or multiple texts (e.g., Bannert, 2003; Azevedo and Cromley, 2004; Naumann et al., 2008; see Cho and Afflerbarch, 2017 for an overview). If, however from a construct perspective metacognitive regulation is one central aspect of reading digital text, it would be counterintuitive to view it as a threat to validity when knowledge of reading strategies governs the adaptive allocation of time on task, and possibly thereby performance on tasks. Rather, as for comprehension skill, the present results evidence one particular mechanism by which knowledge of reading strategies might translate itself into successful reading of digital text.

This notion does not necessarily hold for enjoyment of reading. According to the reasoning put forward in the present research, students high in reading enjoyment do a better job in aligning their time on task behavior with task difficulty because they see reading as less an aversive task. For this reason, it is easier for them than for their peers lower in reading enjoyment to maintain effort and invest time in difficult tasks. Thus, according to the present reasoning, the positive association of reading enjoyment, or reading motivation in general, and reading skill, does not only arise because students higher in reading enjoyment, or motivation, come from higher SES backgrounds, from where they also can acquire better skill (e.g., Artelt et al., 2010). Also, it is not (only) that higher enjoyment or motivation longitudinally bring about better skills, or the reverse (e.g., Becker et al., 2010; Retelsdorf et al., 2011). Rather, just like comprehension skill and knowledge of reading strategies, reading enjoyment seems to be among the variables that govern the process of task engagement in the assessment situation itself and thereby may bring about better task performance and thus a higher level of estimated skill.

Other than comprehension skill and knowledge of reading strategies however, reading enjoyment is not necessarily to be seen as a part of the construct "skill in reading digital text". In other words: A skilled digital reader, who is not in command of comprehension skills is as self-contradictory an idea as a skilled digital reader, who is not in possession of knowledge of reading strategies. In contrast to this, a skilled digital reader who simply does not enjoy reading might be a rare observation, as reading skill and enjoyment are usually positively correlated. The notion of such a reader, however, is not at all a contradictory idea.

From these perspectives, practical implications for the design of assessments, and practical implications for reading in other task-oriented reading situations such as learning are not quite aligned with one another: The finding that reading enjoyment, even if to only a small extent, enhances the adaptive allocation of time might pose a threat to valid interpretations of test scores. On the other hand, it once again highlights the crucial role of motivation in bringing about dedicated and engaged reading behavior, which in turn has been found to be a crucial determinant of learning from text (Guthrie et al., 2012, 2013). This, in turn, once again highlights the need for students to develop motivational traits and attitudes that help them to put in the effort required to cope with difficult and demanding digital texts. Obviously, this notion holds also for knowledge of reading strategies, and, last not least, comprehension skills. Putting students in a position to adequately mobilize cognitive resources when dealing with digital text seems especially important, as digital text to an increased degree requires students not only to "navigate" (see section "The present research" above), but also to evaluate text (Salmerón et al., 2018), a process which is cognitively demanding (Richter and Maier, 2017), and which many students find difficult to perform (e.g., Brante and Strømsø, 2018).

Limitations and Directions

Obviously, the interpretations of the present results put forward here are not without alternative. This is a result of the correlational nature of many large-scale assessment data sets, the present amongst them. This means that there is a host of person-related variables that might, in theory, account for the present results but were unaccounted for in the present research. One candidate here is for example dispositional, or trait self-control, a variable that was found to be related to test-taking effort (Lindner et al., 2017), and thus may very well predict how well students are prepared to align their time on task-behavior to task difficulties. Another variable not taken into account here are specifics of students' preparedness to cope with digital text, such as their navigation skills. Against the background of navigation being a central requirement of reading digital text (Salmerón et al., 2018), students' preparedness to cope with navigation demands might also govern how much time they are prepared to invest in hard, and how little time they might need to complete easy digital reading tasks. Future research thus should seek out additional variables that might affect students' preparedness to adapt their time on task behavior. Analyses such as these might also explain why some lesser skilled readers in fact did align their time on task behavior with task difficulties, while others did not (see **Figure 4**): Perhaps some poorer comprehenders are in possession of other skills than comprehension, which compensate for their lesser comprehension skill, allowing them to nevertheless building an adequate task model. For instance, recent research has shown that problem solving skills interact with comprehension skills in predicting digital reading in such a compensatory fashion (Naumann et al., 2018).

A second limitation comes from the fact that the three predictors used in the present analyses were measured with largely varying numbers of items (although the reliabilities were comparable). Thus, in an assessment using a more comprehensive measurement of reading enjoyment, or knowledge of reading strategies, the interactions of these variables with task difficulty might have been even stronger, maybe at the expense of the interaction between comprehension skill and task difficulty. Future research will have to seek out whether the small effect size for the interaction between reading enjoyment and task difficulty is indeed a function of the comparatively small number of items, or if – to the contrary – after controlling for comprehension skill, there is little variance left to be explained for reading enjoyment due to these variables being positively correlated (see **Table 1** and **Supplementary Material 1**).

In a similar vein, future research should overcome not only the limited number of items, but also the limited operationalization of reading motivation used in the present research. For example, in real-life task oriented reading it might well be the case that topic interest is even more important than reading enjoyment in shaping the interaction between task difficulty and time on task: It might well be that a person who only moderately enjoys reading (and might even be a modest comprehender) will invest time even in a hard task if they have a very high interest in the topic. Research such as this however must be left to future experiments, as large-scale reading assessments usually cannot provide data on topic interest due to the variety of topics addressed by the texts in the assessment. Finally, future research into the role of motivational variables might consider not only linear (as in the present research), but also more complex non-linear effects. A motivated reader for example, who however is in possession of only moderate comprehension skills, might adapt time on task behavior to task difficulty in a non-linear fashion. Such a reader might invest time especially in moderately difficult tasks, while realizing that very hard tasks are beyond their skill level.

A third limitation, and possible avenue for future research, comes from the fact that only one domain was investigated in this research. Future studies might look at how e.g., the time on task behavior in mathematics might be shaped by students' mathematical skills. For example, the ability to "formulate" a mathematical problem, i.e., to "translate from a real-world setting to the domain of mathematics and provide the real-world problem with mathematical structure, representations, and specificity" (OECD, 2013, p. 28) might be conceptually related to building an adequate task model in a reading task. Also, subjective interest in mathematics might moderate the task difficulty-time on task relationship in a fashion similar to the respective effects of reading enjoyment that were found in the present research. With reference to tasks, requirements in the present research were operationalized as the tasks' overall difficulties, as estimated by the international calibration of the Digital Reading Assessment items (OECD, 2012). Building on the present results, future research might seek out which specific features of a digital reading task that might make it "hard" (on the word, sentence, text, or intertextual level) in particular drives time on task behavior in conjunction with person level variables such as the ones addressed here. From an analysis such as this, the question might also be addressed how digital reading assessment tasks might be constructed in a way that variables such as reading enjoyment, or other person level variables that are not part of the targeted construct, do not interact with task features in bringing about task engagement processes that presumably impact task performance and thus estimated abilities. With large scale assessments such as TIMSS or PISA moving toward being computer-based in general (Mullis, 2017; OECD, 2017), analyses such as these could be carried out routinely as part of field trials, and thereby potentially increase the validity of the assessments and in turn the veridicality of conclusions drawn for educational policy and practice.

ETHICS STATEMENT

This study is based on a secondary analysis of OECD PISA 2009 data. Data collection was in accordance with APA ethical standards. The author had OECD's permission to utilize the PISA

2009 Digital Reading Data log files, where not publicly available, for scholarly research and thanks the OECD for this permission.

AUTHOR CONTRIBUTIONS

JN helped in conceiving the assessment materials, conceived and conducted the analyses, and wrote the manuscript.

FUNDING

This study was supported by the German Ministry of Education and Science (BMBF), Grant 01LSA1504A.

ACKNOWLEDGMENTS

I would like to thank the BMBF for their support. I would like to thank the OECD for permission to use the PISA 2009 Digital Reading Logfile Data for scholarly work. I would also like to thank Mr. Adrian Peach for careful proofreading of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01429/full#supplementary-material

REFERENCES

Adams, R. J., Wu, M. L., and Wilson, M. (2012). The rasch rating model and the disordered threshold controversy. *Educ. Psychol. Meas.* 72, 547–573. doi: 10.1177/0013164411432166

AERA, APA, and NCME (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Aiken, L. S., West, S. G., and Reno, R. R. (2003). *Multiple Regression: Testing and Interpreting Interactions* ([Nachdr.]). Newbury Park, CA: SAGE.

Artelt, C., Beinicke, A., Schlagmüller, M., and Schneider, W. (2009). Diagnose von Strategiewissen beim Textverstehen [Assessing knowledge about reading strategies]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie* 41, 96–103. doi: 10.1026/0049-8637.41.2.96

Artelt, C., Naumann, J., and Schneider, W. (2010). "Lesemotivation und Lernstrategien [Reading motivation and learning strategies]," in *PISA 2009: Bilanz nach Einem Jahrzehnt*, eds E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, et al. (Münster: Waxmann), 73–112.

Azevedo, R., and Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *J. Educ. Psychol.* 96, 523–535. doi: 10.1037/0022-0663.96.3.523

Bannert, M. (2003). Effekte metakognitiver Lernhilfen auf den Wissenserwerb in vernetzten Lernumgebungen [Effects of metacognitive help on knowledge acquisition in web-based learning environments]. *Zeitschrift für Pädagogische Psychologie* 17, 13–25. doi: 10.1024//1010-0652.17.1.13

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Becker, M., McElvany, N., and Kortenbruck, M. (2010). Intrinsic and extrinsic reading motivation as predictors of reading literacy: a longitudinal study. *J. Educ. Psychol.* 102, 773–785. doi: 10.1037/a0020084

Borgonovi, F., and Biecek, P. (2016). An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test. *Learn. Indiv. Differ.* 49, 128–137. doi: 10.1016/j.lindif.2016.06.001

Borsboom, D., Mellenbergh, G. J., and van Heerden, J. (2003). The theoretical status of latent variables. *Psychol. Rev.* 110, 203–219. doi: 10.1037/0033-295X.110.2.203

Brante, E. W., and Strømsø, H. I. (2018). Sourcing in text comprehension: a review of interventions targeting sourcing skills. *Educ. Psychol. Rev.* 30, 773–799. doi: 10.1007/s10648-017-9421-7

Britt, M. A., Rouet, J.-F., and Durik, A. (2018). "Representations and processes in multiple source use," in *Educational Psychology Handbook Series. Handbook of Multiple Source Use*. eds J. L. G. Braasch, I. Bråten, and M. T. McCrudden (London: Routledge), 17–33.

Cerdán, R., Gilabert, R., and Vidal-Abarca, E. (2011). Selecting information to answer questions: strategic individual differences when searching texts. *Learn. Indiv. Differ.* 21, 201–205. doi: 10.1016/j.lindif.2010.11.007

Cho, B.-Y., and Afflerbarch, P. (2017). "An evolving perspective of constructively responsive reading comprehension strategies in multilayered digital text environments," in *Handbook of Research on Reading Comprehension*. eds S. E. Israel and G. G. Duffy (New York, NY: The Guilford Press), 109–134.

Cho, B.-Y., Afflerbarch, P., and Han, H. (2018). "Strategic processing in accessing, comprehending, and using multiple sources online," in *Educational Psychology Handbook Series. Handbook of Multiple Source Use*. eds J. L. G. Braasch, I. Bråten, and M. T. McCrudden (London: Routledge), 133–150.

Coiro, J. (2011). Predicting reading comprehension on the internet. *J. Liter. Res.* 43, 352–392. doi: 10.1177/1086296X11421979

De Boeck, P. (2008). Random item IRT models. *Psychometrika* 73, 533–559. doi: 10.1007/S11336-008-9092-X

Debeer, D., Buchholz, J., Hartig, J., and Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *J. Educ. Behav. Stat.* 39, 502–523. doi: 10.3102/1076998614558485

Debeer, D., and Janssen, R. (2013). Modeling item-position effects within an IRT framework. *J. Educ. Meas.* 50, 164–185. doi: 10.1111/jedm.12009

Dufresne, A., and Kobasigawa, A. (1989). Children's spontaneous allocation of study time: differential and sufficient aspects. *J. Exp. Child Psychol.* 47, 274–296. doi: 10.1016/0022-0965(89)90033-7

Gelman, A., and Su, Y.-S. (2016). *Arm: Data Analysis using Regression and Multilevel/Hierarchical Models. R package version 1.9-3*.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., and Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: insights from a computer-based large-scale assessment. *J. Educ. Psychol.* 106, 608–626. doi: 10.1037/a0034716

Greiff, S., Wüstenberg, S., and Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Comp. Educ.* 91, 92–105. doi: 10.1016/j.compedu.2015.10.018

Guthrie, J. T., Klauda, S. L., and Ho, A. N. (2013). Modeling the relationships among reading instruction, motivation, engagement, and achievement for adolescents. *Read. Res. Q.* 48, 9–26. doi: 10.1002/rrq.035

Guthrie, J. T., Wigfield, A., and You, W. (2012). "Instructional contexts for engagement and achievement in reading," in *Handbook of Research on Student Engagement*. eds S. L. Christenson, A. L. Reschly, and C. Wylie (Boston, MA: Springer), 601–634.

Hahnel, C., Goldhammer, F., Kröhne, U., and Naumann, J. (2018). The role of reading skills in the evaluation of online information gathered from search engine environments. *Comput. Hum. Behav.* 78, 223–234. doi: 10.1016/j.chb.2017.10.004

Hedges, L. V., and Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychol. Methods* 3, 486–504. doi: 10.1037/1082-989X.3.4.486

Inzlicht, M., and Schmeichel, B. J. (2012). What is ego depletion? Toward a mechanistic revision of the resource model of self-control. *Perspect. Psychol. Sci.* 7, 450–463. doi: 10.1177/1745691612454134

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge: Cambridge University Press.

Lawless, K., and Schrader, P. G. (2008). "Where do we go now?," in *Handbook of Research on New literacies*. eds J. Coiro, M. Knobel, C. Lankshear, and D. J. Leu (Abingdon: Routledge), 267–296.

Lindner, C., Nagy, G., Ramos Arhuis, W. A., and Retelsdorf, J. (2017). A new perspective on the interplay between self-control and cognitive performance:

modeling progressive depletion patterns. *PloS One* 12:e0180149. doi: 10.1371/journal.pone.0180149

Lindner, C., Nagy, G., and Retelsdorf, J. (2018). The need for self-control in achievement tests: changes in students' state self-control capacity and effort investment. *Soc. Psychol. Educ.* 21, 1113–1131. doi: 10.1007/s11218-018-9455-9

Lockl, K., and Schneider, W. (2002). Zur Entwicklung des selbstregulierten Lernens im Grundschulalter: zusammenhänge zwischen Aufgabenschwierigkeit und Lernzeiteinteilung [The development of self-regulated lin elementary school children: associations between task difficulty and allocation of study time]. *Psychologie in Erziehung und Unterricht* 49, 3–16.

Lockl, K., and Schneider, W. (2003). Metakognitive Überwachungs- und Selbstkontrollprozesse bei der Lernzeiteinteilung von Kindern [Metacognitive monitoring and self-control processes for children's allocation of study time]. *Zeitschrift für Pädagogische Psychologie* 17, 173–183. doi: 10.1024//1010-0652.17.34.173

Mañá, A., Vidal-Abarca, E., and Salmerón, L. (2017). Effect of delay on search decisions in a task-oriented reading environment. *Metacogn. Learn.* 12, 113–130. doi: 10.1007/s11409-016-9162-x

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika* 47, 149–174. doi: 10.1007/BF02296272

Mullis, I. V. S. (2017). "Introduction," in *TIMSS 2019 Assessment Frameworks*. eds I. V. S. Mullis and M. O. Martin (Boston, MA: Boston College), 1–10.

Nagy, G., Nagengast, B., Becker, M., Rose, N., and Frey, A. (2018a). Item position effects in a reading comprehension test: an IRT study of individual differences and individual correlates. *Psychol. Test Assess. Model.* 60, 165–187.

Nagy, G., Nagengast, B., Frey, A., Becker, M., and Rose, N. (2018b). A multilevel study of position effects in PISA achievement tests: student- and school-level predictors in the German tracked school system. *Assess. Educ.* 55, 1–22. doi: 10.1080/0969594X.2018.1449100

Naumann, J. (2015). A model of online reading engagement: linking engagement, navigation, and performance in digital reading. *Comput. Hum. Behav.* 53, 263–277. doi: 10.1016/j.chb.2015.06.051

Naumann, J., and Goldhammer, F. (2017). Time-on-task effects in digital reading are non-linear and moderated by persons' skills and tasks' demands. *Learn. Indiv. Differ.* 53, 1–16. doi: 10.1016/j.lindif.2016.10.002

Naumann, J., Pucite, L., Goldhammer, F., Greiff, S., and Eichmann, B. (2018). *Interactive Effects of Comprehension and Problem Solving Skills on Digital Reading Performance and Navigation*. New York, NY: American Educational Research Association.

Naumann, J., Richter, T., Christmann, U., and Groeben, N. (2008). Working memory capacity and reading skill moderate the effectiveness of strategy training in learning from hypertext. *Learn. Indiv. Differ.* 18, 197–213. doi: 10.1016/j.lindif.2007.08.007

Naumann, J., and Salmerón, L. (2016). Does navigation always predict performance? Effects of navigation on digital reading are moderated by comprehension skills. *Int. Rev. Res. Open Distrib. Lear.* 17, 42–59. doi: 10.19173/irrodl.v17i1.2113

OECD (2009). *PISA 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science*. Paris: OECD Publishing.

OECD (2010). *PISA 2009 Results: What Students Know and can do*. Paris: OECD Publishing.

OECD (2011). *PISA 2009 Results vol. VI: Students on Line*. Paris: OECD Publishing.

OECD (2012). *PISA 2009 Technical Report*. Paris: OECD Publishing.

OECD (2013). *PISA 2012 Assessment and Analytical Framework*. Paris: OECD Publishing.

OECD (2015). *Students, Computers and Learning: Making the Connection*. Paris: OECD Publishing.

OECD (2017). *PISA 2015 Technical Report*. Paris: OECD Publishing.

Pressley, M., Borkowski, J. G., and Schneider, W. (1989). Good information processing: what it is and how education can promote it. *Int. J. Educ. Res.* 13, 857–867. doi: 10.1016/0883-0355(89)90069-4

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Retelsdorf, J., Köller, O., and Möller, J. (2011). On the effects of motivation on reading performance growth in secondary school. *Learn. Instruct.* 21, 550–559. doi: 10.1016/j.learninstruc.2010.11.001

Richter, T., and Maier, J. (2017). Comprehension of multiple documents with conflicting information: a two-step model of validation. *Educ. Psychol.* 52, 148–166. doi: 10.1080/00461520.2017.1322968

Rouet, J.-F., and Britt, M. A. (2011). "Relevance processes in multiple document comprehension," in *Text Relevance and Learning from Text*. eds G. J. Schraw, J. P. Magliano, and M. T. McCrudden (Charlotte, NC: Information Age Pub), 19–52.

Rouet, J.-F., Britt, M. A., and Durik, A. M. (2017). RESOLV: readers' representation of reading contexts and tasks. *Educ. Psychol.* 52, 200–215. doi: 10.1080/00461520.2017.1329015

Rouet, J.-F., and Le Bigot, L. (2007). Effects of academic training on metatextual knowledge and hypertext navigation. *Metacogn. Learn.* 2, 157–168. doi: 10.1007/s11409-007-9011-z

Rouet, J.-F., Ros, C., Goumi, A., Macedo-Rouet, M., and Dinet, J. (2011). The influence of surface and deep cues on primary and secondary school students' assessment of relevance in Web menus. *Learn. Instruc.* 21, 205–219. doi: 10.1016/j.learninstruc.2010.02.007

Salmerón, L., Cerdán, R., and Naumann, J. (2015a). How adolescents navigate Wikipedia to answer questions/¿Cómo navegan los adolescentes en Wikipedia para contestar preguntas? *Infancia y Aprendizaje* 38, 435–471. doi: 10.1080/02103702.2015.1016750

Salmerón, L., Vidal-Abarca, E., Martínez, T., Mañá, A., Gil, L., and Naumann, J. (2015b). Strategic decisions in task-oriented reading. *Span. J. Psychol.* 18:E102. doi: 10.1017/sjp.2015.101

Salmerón, L., Strømsø, H. I., Kammerer, Y., Stadtler, M., and van den Broek, P. (2018). "Comprehension processes in digital reading," in *Learning to Read in a Digital World*. eds M. Barzillai, J. Thomson, S. Schroeder, and P. van den Broek (Amsterdam: John Benjamins Publishing Company), 91–120.

Schneider, W., and Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychol. Rev.* 84, 1–66. doi: 10.1037/0033-295X.84.1.1

Serrano, M.-Á., Vidal-Abarca, E., and Ferrer, A. (2018). Teaching self-regulation strategies via an intelligent tutoring system (TuinLECweb): effects for low-skilled comprehenders. *J. Comp. Assist. Learn.* 34, 515–525. doi: 10.1111/jcal.12256

Shiffrin, R. M., and Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychol. Rev.* 84, 127–190. doi: 10.1037/0033-295X.84.2.127

Son, L. K., and Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *J. Exp. Psychol.* 26, 204–221. doi: 10.1037/0278-7393.26.1.204

Sternberg, R. J. (1986). Inside intelligence: cognitive science enables us to go beyond intelligence tests and understand how the human mind solves problems. *Am. Sci.* 74, 137–143.

Vidal-Abarca, E., Mañá, A., and Gil, L. (2010). Individual differences for self-regulating task-oriented reading activities. *J. Educ. Psychol.* 102, 817–826. doi: 10.1037/a0020062

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* 36, 1–48.

Weinstein, C. E., and Mayer, R. E. (1986). "The teaching of learning strategies," in *Handbook of Research on Teaching: A Project of the American Educational Research Association*, 3rd Edn. ed. M. C. Wittrock (New York, NY: Macmillan), 315–332.

Weirich, S., Hecht, M., Penk, C., Roppelt, A., and Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Appl. Psychol. Meas.* 41, 115–129. doi: 10.1177/0146621616676791

Winne, P. H., and Hadwin, A. E. (1998). "Studying as self-regulated learning," in *Educational Psychology Series. Metacognition in Educational Theory and Practice*, eds A. C. Graesser, D. J. Hacker, and J. Dunlosky (Mahwah, N.J: L. Erlbaum Associates), 277–304.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 28 June 2019
doi: 10.3389/fpsyg.2019.01313

Fine-Grained Assessment of Children's Text Comprehension Skills

Marije den Ouden[1,2], *Jos Keuning*[1] *and Theo Eggen*[1,2]*

[1]Cito, Arnhem, Netherlands, [2]Faculty of Behavioural, Management and Social Sciences, University of Twente, Enschede, Netherlands

Text comprehension is an essential skill for achievement in personal, academic, and professional life. Therefore, it is tremendously important that children's text comprehension skills are actively monitored from an early stage. Text comprehension is, however, a complex process in which different reading abilities continuously interact with each other on the word, sentence, and text levels. In educational practice, various tests are used to measure these different reading abilities in isolation, which makes it very difficult to understand why a child scores high or low on a specific reading test and to adequately tailor reading instruction to the child's needs. Dynamic assessment has the potential to offer insights and guidance to teachers as cognitive processes that are important for learning are examined. In dynamic tests, students receive mediation through instruction when answering test questions. Although computer-based dynamic assessment in the reading domain holds potential, there is almost no support for the validity of dynamic measures of text comprehension. The aim of the present study is to determine design principles for the intended use of computer-based dynamic assessment of text comprehension. Based on the dynamic assessment literature, we developed a model for assessing the different reading abilities in conjunction. The assumption is that this model gives a fine-grained view of children's strengths and weaknesses in text comprehension and provides detailed information on children's instructional needs. The model was applied in a computer-based (fourth-grade) reading assessment and evaluated in practice through a three-group experimental design. We examined whether it is possible to (1) measure different aspects of the reading process in conjunction in order to obtain a full understanding of children's text comprehension skills, (2) measure children's learning potential in text comprehension, and (3) provide information on their instructional needs. The results show that while the model helped in explaining the children's text comprehension scores, unexpectedly, mediation did not clearly lead to progress in text comprehension. Based on the outcomes, we substantiate design principles for computer-based dynamic assessment of text comprehension.

Keywords: computer-based assessment, design principles, dynamic assessment, instructional needs, learning potential, reading process, text comprehension

OPEN ACCESS

Edited by:
Frank Goldhammer,
German Institute for International Educational Research (LG), Germany

Reviewed by:
Maj-Britt Isberner,
University of Kassel, Germany
Tobias Doerfler,
Heidelberg University of Education, Germany

****Correspondence:***
Theo Eggen
theo.eggen@cito.nl

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *23 October 2018*
Accepted: *20 May 2019*
Published: *28 June 2019*

Citation:
den Ouden M, Keuning J and Eggen T (2019) Fine-Grained Assessment of Children's Text Comprehension Skills. Front. Psychol. 10:1313. doi: 10.3389/fpsyg.2019.01313

INTRODUCTION

Text comprehension is an important skill for personal fulfillment and for achieving academic and professional success. Nevertheless, it is also a very complex skill involving different cognitive abilities that interact on different levels. At lower levels, word identification skills and knowledge of word meanings are essential for understanding text (Perfetti and Hart, 2001; Perfetti, 2017). At higher levels, text comprehension is influenced by the ability to make inferences or monitor comprehension (Perfetti et al., 2005). This complex nature results in a variety of possible causes underlying problems encountered in text comprehension (Cain and Oakhill, 2006; Colenbrander et al., 2016; Kleinsz et al., 2017). Whereas most primary school teachers underline the importance of developing good text comprehension skills, they also point to difficulties understanding the reading problems children encounter. We aim to develop a framework for fine-grained assessment of text comprehension skills that supports teachers in understanding children's text comprehension problems.

Measuring Text Comprehension Skills

Research on text comprehension has advanced a number of theories on the different parts of the reading process. Due to the complex nature of text comprehension, interactive models of the reading process arguably provide the best framework for understanding and studying this concept (Stanovich, 1980; Perfetti, 1999; Cain et al., 2017). These models have in common that they describe the reading process of interaction on different levels (e.g., word, sentence, and text levels) and often make a distinction between processing information explicitly stated in the text and deriving information implicitly stated in the text. One of the most influential models is the construction-integration model (Van Dijk and Kintsch, 1983; Kintsch, 1988, 1998), which describes the reciprocal relation between the *construction* of a text-based model and its *integration* into a situation model. A distinction is made between combining all information that is explicitly stated in the text on the word, sentence, and text levels (text model) and interpreting this information, together with prior knowledge, as a coherent whole (situation model). Verhoeven and Perfetti (2008) place greater emphasis on the role of word knowledge by conceptualizing text comprehension as an interaction between word identification and word-to-text integration (see **Figure 1**). Words are identified by combining orthographic, phonological, and semantic representations. The quality of these representations significantly influences text comprehension (Perfetti and Hart, 2001). Identified words can be linked to each other in order to give meaning to a sentence, and sentences can be linked through inferences based on explicit (text model) and implicit (situation model) information.

In order to obtain a full understanding of children's text comprehension skills, educational assessment should cover the various aspects of the reading process. In educational practice,

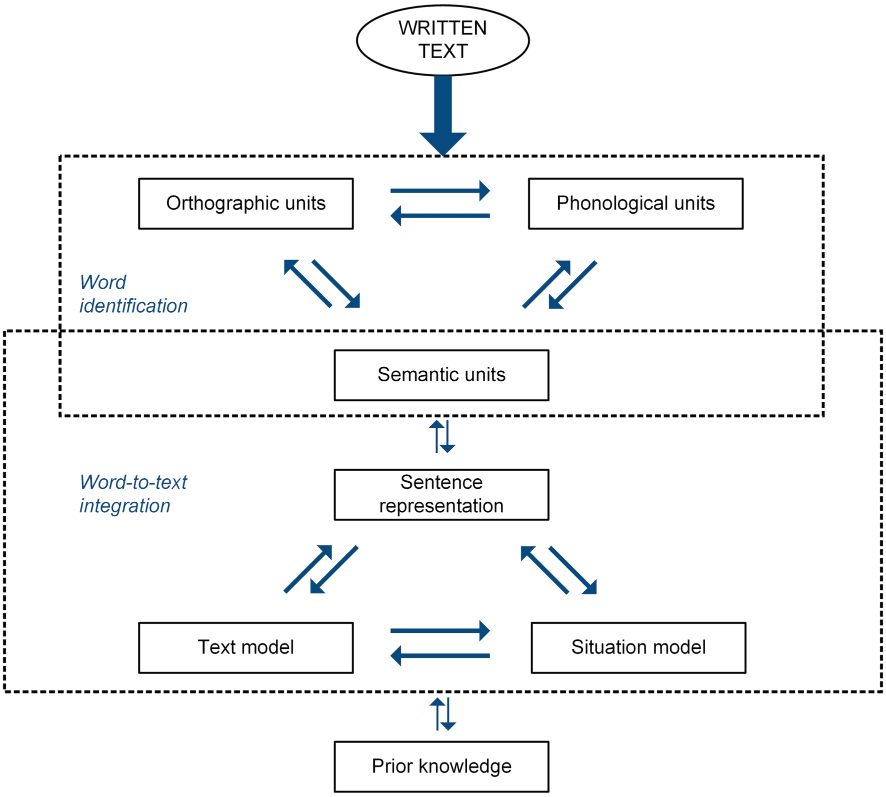


FIGURE 1 | Process of text comprehension, as modeled by Verhoeven and Perfetti (2008).

a variety of tests are used to measure these different aspects. For example, nationally standardized tests (NSTs) are deployed for student monitoring, i.e., monitoring students' progress on skills as text comprehension, vocabulary, and word decoding. Additionally, tests originating from teaching materials are administered to evaluate knowledge acquired through education. Consequently, different aspects of the reading process are evaluated through different reading tests, and even tests that are supposed to measure the same construct show only modest intercorrelation (Nation and Snowling, 1997; Keenan et al., 2008). This fragmentary way of measuring reading ability problematizes the interpretation of the test results in a coherent way. Therefore, this way of measuring makes it very difficult to understand why a child scores high or low on a specific reading test and to adequately tailor reading instruction to the child's needs. Moreover, measuring different aspects of the reading process in isolation is questionable in terms of its interactive nature. It might also be difficult to eliminate every aspect other than that intended for measurement. For example, poor vocabulary can result in an underestimation of inference-making skills (Segers and Verhoeven, 2016; Daugaard et al., 2017; Swart et al., 2017). These issues could be addressed by measuring text comprehension in a more comprehensive way, i.e., measuring different aspects of the reading process in conjunction.

Furthermore, commonly used tests usually provide insufficient diagnostic information, e.g., information about students' misconceptions and learning potential (Fani and Rashtchi, 2015). Thus, these tests provide only little support for teachers in aligning their reading instruction to the educational needs of their students. *Dynamic assessment* has the potential to offer insights and guidance to teachers as cognitive processes that are important for learning are examined (Lidz and Elliott, 2000; Elliott, 2003). In dynamic tests, students receive mediation through instruction when answering test questions. Dynamic assessment of text comprehension skills can provide teachers with information to identify students' capabilities as well as their specific needs for training in the reading domain (Dörfler et al., 2017).

Dynamic Assessment

The idea of dynamic assessment is based on Vygotsky's (1978) theory of the zone of proximal development (ZPD), wherein human abilities are perceived in a constant state of flux and are sensitive to sources of mediation that can feed learning mechanisms. Lantolf and Poehner (2007) describe two approaches to dynamic assessment: the interactionist and interventionist approaches. The interactionist approach involves the traditional dynamic assessments, whereby the type and amount of instruction provided depend on one-on-one interaction between the teacher and student. The instruction is completely attuned to the responsiveness of the student (Lantolf and Poehner, 2007). In the interactionist approach, the goal is to reach the maximum performance for each individual student. By contrast, the interventionist approach involves standardized instruction that is arranged in advance and quantified during the assessment. This approach focuses on determining the amount and nature of instruction a student needs in order to reach a pre-specified performance level. An interventionist dynamic assessment is less time-consuming, and its results are more comparable across students, since every student is tested according to the same procedure. It enhances efficiency in terms of the number of students that can be tested simultaneously, especially when the assessment is digitalized (Poehner and Lantolf, 2013).

Computer-based interventionist dynamic assessment can be elaborated through different designs. Sternberg and Grigorenko (2002) distinguish between the sandwich and cake designs. The sandwich design can be defined as a test-train-test design in which a pretest is followed by some intervention or instruction (see **Figure 2**), and a posttest comparable to the pretest is subsequently administered to all students. With this design, one can determine the extent to which students are able to improve when instruction is offered (Tzuriel, 2000). Performance before and after this instruction can be compared in order to examine students' ZPD or their potential to learn. The cake design can be defined as a train-within-test design in which instruction follows immediately after an incorrect response to an item (see **Figure 2**). The instruction can be presented as a graded series of instructional hints that guide the student toward the correct response, referred to as the graduated prompts approach (Brown and Ferrara, 1985; Campione and Brown, 1987). This approach determines the amount of aid a student needs to solve the problem (Tzuriel, 2000). The number of hints needed to find the correct response is often used as an indication of students' ZPD or learning potential.

Model for Fine-Grained Assessment

Measuring text comprehension in a comprehensive and dynamic way given the discussed purpose holds some challenges. First, a test of this nature should provide a full understanding of students' text comprehension skills. All parts and interactions of the reading process should ideally be examined in conjunction. Using the model of Verhoeven and Perfetti (2008), these can be summarized in the constructs word-form

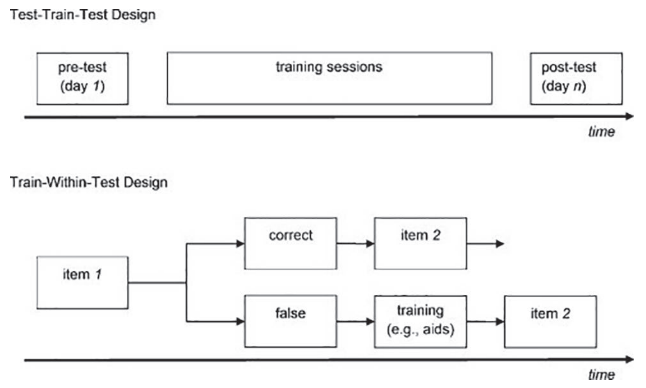


FIGURE 2 | Traditional dynamic assessment design. A test-train-test design (above) with pre-, posttest, and separate training sessions and a train-within-test design (below) with only one session, including training parts (Dörfler et al., 2009).

knowledge (orthographic and phonological representations), word-meaning knowledge (semantic representations), local cohesion inferences (word-to-text integration), and global understanding (text and situation model). Moreover, this test should inform teachers about the educational needs of their students as well as of the efficacy of intervention. Furthermore, the administration of this test should be feasible. It should take a limited amount of time and ought to be clearly beneficial to both the teacher and student. All these considerations have been accounted for in the assessment model presented in **Figure 3**, which presents an amalgamation of the sandwich and cake designs.

Both the sandwich and cake designs show some difficulties when used for dynamic reading assessment. In the cake design, quantifying the amount of instruction students need in order to find correct responses can provide an indication of their learning potential; however, it does not allow for modeling the effect of instruction. In the sandwich design, change in performance level caused by the training can be modeled. However, this overall effect cannot be linked to specific types of instruction, since there is only one intervention phase. Multiple training sessions and posttests can address this issue but would nonetheless be highly time-consuming. By combining the sandwich and cake designs, the overall effect of instruction (i.e., learning potential) can be determined and can also be linked to the amount and nature of the instruction offered.

Following the proposed assessment model, a test with the same set of items measuring global understanding is presented in two respective measurement occasions. At the first measurement occasion, a set of items is presented without instruction, and at the second measurement occasion, a set of items is presented with item-level instructions. The instructions consist of several supportive scaffolding questions related to word-form knowledge, word-meaning knowledge, and local cohesion inferences, along with corresponding feedback. At the second measurement occasion, children are thus trained in successfully completing the global text comprehension task by first teaching them the necessary knowledge at word- and sentence level.

Scaffolding and Feedback

As discussed earlier, dynamic assessment is characterized by the inclusion of instruction during test administration. In this way, dynamic tests provide information about the educational needs of students as well as possible intervention. Research has showed that students with similar initial abilities can benefit differentially from instruction (Tzuriel, 2000). Moreover, different shortcomings in the process of reading might require different approaches from teachers in providing guidance and instruction (Fuchs et al., 2012). In the proposed design, the instruction phase consists of scaffolding questions and feedback on the responses to these questions.

Scaffolding can be defined as providing cognitive support by breaking down tasks into smaller, more manageable parts that are within the student's understanding (Dennen and Burner, 2008). In the case of reading comprehension, determining the main idea of the text is a cognitively demanding task that can be broken down into smaller tasks as determining (the meaning of) important words and making required inferences between sentences or paragraphs. According to Vygotsky's (1978) theory of ZPD, students can achieve their potential level of development if scaffolding is applied to them (Magno, 2010), which can be applied in the form of questions, as recommended by Feng (2009). By using a series of scaffolding questions that focus on different cognitive abilities on different levels of the reading process, we can gradually guide a student toward global understanding of a text. Also, we can determine the extent to which a student is capable of making necessary intermediate steps for gaining global understanding of the text. Moreover, different aspects of the reading process can be measured in this way.

Feedback on students' responses to scaffolding questions is essential for letting them acquire the intended knowledge. Item-based feedback can be presented as either verification or elaboration.

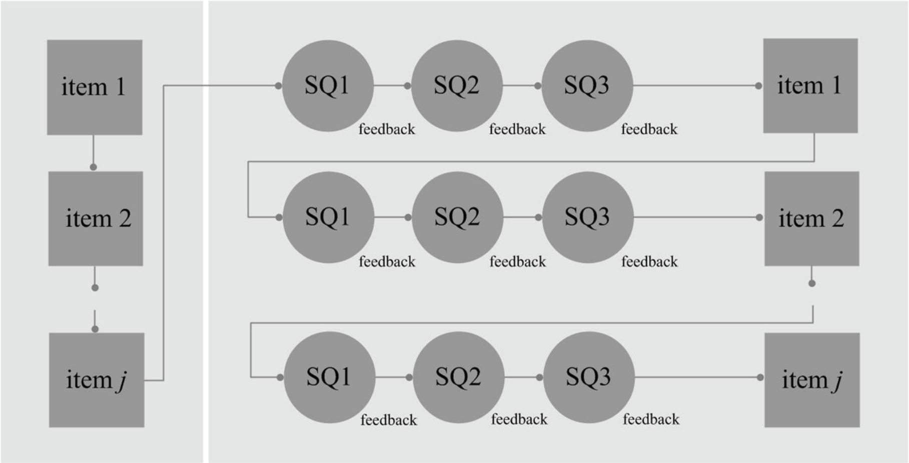


FIGURE 3 | Assessment model for dynamic reading assessment. Performance on the items of the posttest (right squares) can be compared to performance on the items of the pretest (left squares). The change in performance can then be linked to/explained by performance on the scaffolding questions (SQ).

Elaborated feedback is more effective than verification; however, they are most effective when combined (Dörfler et al., 2009; Van der Kleij et al., 2015). Verification feedback simply consists of a confirmation of an (in)correct response. Elaborated feedback could contain error-specific explanations and solution-oriented prompts or could address meta-cognitive processes. In the proposed design, standardized solution-oriented prompts are preferable, since non-contingent feedback has been shown to be more predictive of future achievement than contingent feedback in dynamic assessment (Caffrey et al., 2008).

The Present Study

Although computer-based dynamic assessment in the reading domain holds potential, there are only a few approaches to dynamic assessment available, and thus, there is almost no support for the validity of dynamic measures of text comprehension (Dörfler et al., 2017). The aim of the present study is to determine design principles for the intended use of computer-based dynamic assessment of text comprehension. The proposed, theoretically based assessment model was applied in a computer-based dynamic assessment for text comprehension and tested and evaluated in practice. We examined whether it is possible to (1) measure different aspects of the reading process in conjunction in order to obtain a full understanding of children's text comprehension skills, (2) measure children's learning potential in text comprehension, and (3) provide information on their instructional needs. Learning potential was defined as the difference between two measurements occasions, one in which a global understanding task was administered without scaffolding and one in which the same task was administered in combination with several supportive scaffolding questions related to word-form knowledge, word-meaning knowledge, and local cohesion inferences. In this study, learning potential thus reflected the child's ability to use the help they get in completing the global understanding task. Instructional needs referred to the children's performance on the different scaffolding questions. Failure on one specific subskill implied that there was a need for additional instruction on that subskill. Based on the conclusions, we substantiate design principles for computer-based dynamic assessment of text comprehension.

MATERIALS AND METHODS

Participants

The study was conducted in cooperation with a school consortium of which four schools participated with their fourth-grade students. Three schools participated with one school class, and one school participated with two school classes. The schools were located in neighborhoods with average and above-average scores in income, employment, and education level, in comparison with the national standard (The Netherlands Institute for Social Research, 2016). The pretest was administered to 169 fourth-grade students aged approximately 10–11 years old. From the pre- to posttest, one school class consisting of 29 students dropped out.

Materials

Texts

A total of 80 texts were selected from a database managed by Cito Institute for Educational Measurement. The database contained texts from existing sources, e.g., children's books, informative books, and websites. Texts were evaluated by T-scan, an analysis tool for Dutch texts to assess the complexity of the text (Pander Maat et al., 2014). The selected texts were found to be appropriate in terms of difficulty following an evaluation of different text attributes, e.g., word difficulty, sentence complexity, verbiage, and referential and causal coherence. Both informative and narrative texts were included. The selected texts contained between 112 and 295 words, averaging 205 words.

Tasks

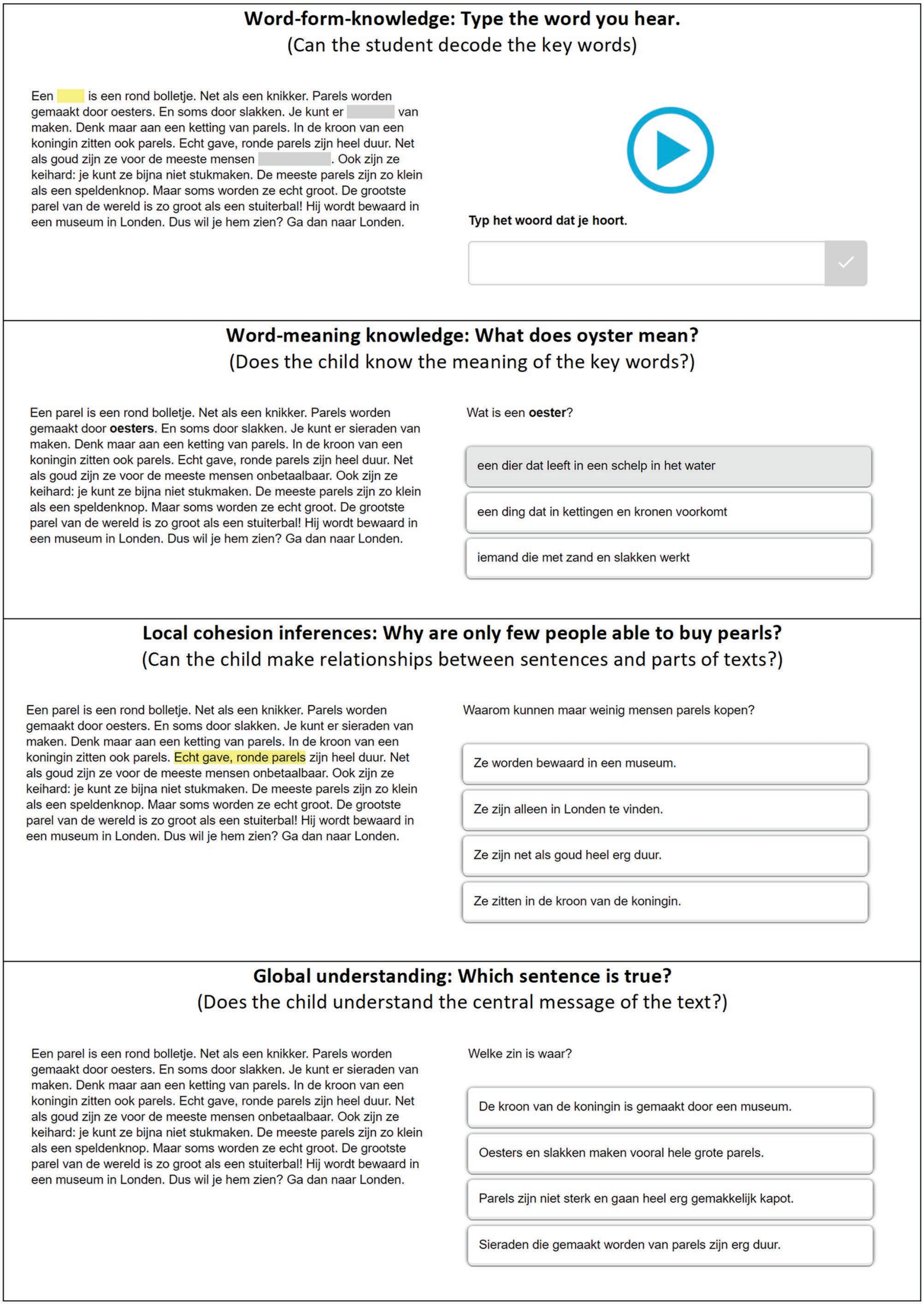
For every text, four tasks were constructed and screened by a group of reading experts. All tasks corresponding to one text were constructed by one reading expert, screened by two other reading experts and, when necessary, adjusted by the first reading expert. The tasks covered different parts of the reading process, as modeled by Verhoeven and Perfetti (2008), as they represented the constructs *word-form knowledge*, *word-meaning knowledge*, *local cohesion inferences*, and *global understanding*.

The different tasks were separately pre-examined in a trial with paper-based tests. Each test consisted of 40 tasks that measured the same construct. Each test was administered to at least two school classes, which resulted in 40–97 administrations per test with a total of 629 administrations. In this preliminary research, all tasks were found to be highly reliable and appropriate with respect to level of difficulty. The resulting item bank consisted of 80 texts and corresponding tasks and was used for the assembly of the final test. Item statistics (i.e., percentage of correct answers and item-total correlation) were used for the test assembly so as to ensure item quality and to maximize task reliability. Texts with too hard (percentage correct < 0.35) or too easy (percentage correct > 0.90) items, or items with a low item-total correlation (<0.20), were not included in the final test.

The final test consisted of 30 texts and was administered twice, as displayed by the squares in **Figure 3**. During a pretest, each text was presented with one task regarding global understanding of the text. During a posttest, each text was presented with up to four tasks; one task regarding global understanding of the text preceded by, depending on the experimental condition, up to three scaffolding questions with feedback. An example of the tasks is shown in **Figure 4**.

Global Understanding

For every text, the students were asked about the main idea of the text in a multiple choice question with four possible choices. This task measured the ability to integrate all the



Word-form-knowledge: Type the word you hear.
(Can the student decode the key words)

Een   is een rond bolletje. Net als een knikker. Parels worden gemaakt door oesters. En soms door slakken. Je kunt er   van maken. Denk maar aan een ketting van parels. In de kroon van een koningin zitten ook parels. Echt gave, ronde parels zijn heel duur. Net als goud zijn ze voor de meeste mensen . Ook zijn ze keihard: je kunt ze bijna niet stukmaken. De meeste parels zijn zo klein als een speldenknop. Maar soms worden ze echt groot. De grootste parel van de wereld is zo groot als een stuiterbal! Hij wordt bewaard in een museum in Londen. Dus wil je hem zien? Ga dan naar Londen.

Typ het woord dat je hoort.

Word-meaning knowledge: What does oyster mean?
(Does the child know the meaning of the key words?)

Een parel is een rond bolletje. Net als een knikker. Parels worden gemaakt door **oesters**. En soms door slakken. Je kunt er sieraden van maken. Denk maar aan een ketting van parels. In de kroon van een koningin zitten ook parels. Echt gave, ronde parels zijn heel duur. Net als goud zijn ze voor de meeste mensen onbetaalbaar. Ook zijn ze keihard: je kunt ze bijna niet stukmaken. De meeste parels zijn zo klein als een speldenknop. Maar soms worden ze echt groot. De grootste parel van de wereld is zo groot als een stuiterbal! Hij wordt bewaard in een museum in Londen. Dus wil je hem zien? Ga dan naar Londen.

Wat is een **oester**?

- een dier dat leeft in een schelp in het water
- een ding dat in kettingen en kronen voorkomt
- iemand die met zand en slakken werkt

Local cohesion inferences: Why are only few people able to buy pearls?
(Can the child make relationships between sentences and parts of texts?)

Een parel is een rond bolletje. Net als een knikker. Parels worden gemaakt door oesters. En soms door slakken. Je kunt er sieraden van maken. Denk maar aan een ketting van parels. In de kroon van een koningin zitten ook parels. Echt gave, ronde parels zijn heel duur. Net als goud zijn ze voor de meeste mensen onbetaalbaar. Ook zijn ze keihard: je kunt ze bijna niet stukmaken. De meeste parels zijn zo klein als een speldenknop. Maar soms worden ze echt groot. De grootste parel van de wereld is zo groot als een stuiterbal! Hij wordt bewaard in een museum in Londen. Dus wil je hem zien? Ga dan naar Londen.

Waarom kunnen maar weinig mensen parels kopen?

- Ze worden bewaard in een museum.
- Ze zijn alleen in Londen te vinden.
- Ze zijn net als goud heel erg duur.
- Ze zitten in de kroon van de koningin.

Global understanding: Which sentence is true?
(Does the child understand the central message of the text?)

Een parel is een rond bolletje. Net als een knikker. Parels worden gemaakt door oesters. En soms door slakken. Je kunt er sieraden van maken. Denk maar aan een ketting van parels. In de kroon van een koningin zitten ook parels. Echt gave, ronde parels zijn heel duur. Net als goud zijn ze voor de meeste mensen onbetaalbaar. Ook zijn ze keihard: je kunt ze bijna niet stukmaken. De meeste parels zijn zo klein als een speldenknop. Maar soms worden ze echt groot. De grootste parel van de wereld is zo groot als een stuiterbal! Hij wordt bewaard in een museum in Londen. Dus wil je hem zien? Ga dan naar Londen.

Welke zin is waar?

- De kroon van de koningin is gemaakt door een museum.
- Oesters en slakken maken vooral hele grote parels.
- Parels zijn niet sterk en gaan heel erg gemakkelijk kapot.
- Sieraden die gemaakt worden van parels zijn erg duur.

FIGURE 4 | Examples of the four tasks regarding (from above to below) word-form knowledge (SQ1), word-meaning knowledge (SQ2), local cohesion inferences (SQ3), and global understanding.

information provided by the text into a situation model. During the pretest, children had to derive this information from the text themselves. During the posttest, children could use the acquired knowledge from the preceded scaffolding and feedback as guidance for finding the correct response.

Word-Form Knowledge (SQ1)

For every text, children were asked to type in three words in three separate open-ended questions. The words were blurred in the text and presented to the students auditory (see upper part of **Figure 4**). As feedback on an incorrect response, the correct word form was shown in the text for 3 s. This task measured the quality of phonological and orthographical representations of words that were essential for understanding the text. By applying scaffolding and feedback on word-form knowledge, the children could get acquainted with the key words of the text.

Word-Meaning Knowledge (SQ2)

For every text, the students were asked for the meaning of two words in two separate multiple choice questions, each with three possible choices of word definitions. The word in question was bolded in the text. The feedback on an incorrect response included a picture of the word in question. This task measured the quality of the semantic representations of words that were essential for understanding the text. By applying scaffolding and feedback on word-meaning knowledge, children received information about the meaning of the key words of the text.

Local Cohesion Inferences (SQ3)

For every text, the students were asked to make an inference, relevant for understanding the main idea of the text, in one multiple choice question with four possible choices. As feedback on an incorrect response, the relevant phrases or sentences were highlighted in yellow in the text. This task measured the ability to integrate text phrases that were essential for understanding the text. By applying scaffolding and feedback on inference-making, the children were encouraged to think about the cohesion of different text parts.

Procedure

The pretest was divided into two subtests, each with 15 texts that were administered on separate occasions on the same day. The posttest was divided into three subtests, each with 10 texts that were administered on separate occasions spread over two consecutive days. All test administrations took place in the classroom, with a duration of 45 min for each occasion. The posttest was administered 4 weeks after the pretest.

All groups received the same pretest. For the posttest, all students were randomly assigned, within the school classes, to one of three conditions. The first experimental condition (n = 47) received the posttest that included all three different types of scaffolding and feedback for every text, SQ1, SQ2, and SQ3. The second experimental condition (n = 48) received the posttest that included two different types of scaffolding and feedback for every text, SQ1 and SQ2. The control condition received the posttest that included no scaffolding or feedback (n = 45). To ensure active processing of feedback, the students had a second attempt at the scaffolding questions following an incorrect response.

Statistical Analyses

In order to determine to which extent we were able to measure different aspects of the reading process in conjunction, the psychometric quality (i.e., reliability and validity) of the developed test was investigated. Classical test and item analyses were conducted for all scales. Internal consistency was assessed with Cronbach's alpha (α), a lower-bound estimate of reliability, with a value of ≥0.80 indicating good reliability, a value of ≥0.70 indicating sufficient reliability, and a value of <0.70 indicating insufficient reliability (Evers et al., 2010).

Furthermore, construct validity was evaluated through the analysis of a multitrait-multimethod matrix (MTMM; Campbell and Fiske, 1959). For this MTMM, scores on the dynamic assessment scales were linked to previously obtained scores on NSTs for text comprehension, vocabulary, orthography, and math. These tests were administered 4 months earlier with the purpose of monitoring students' progress through primary school. Pearson correlation (r) between the scores on the subscales of the dynamic assessment and NSTs was computed and interpreted as high when $r \geq 0.50$, moderate when $r \geq 0.30$, and low when $r < 0.30$ (Cohen, 1988).

In order to determine to which extent we were able to measure children's learning potential in text comprehension and to provide information on their instructional needs, we investigated learning potential and the effect of scaffolding and feedback on *global understanding*. First, the experimental conditions were compared to the control condition on the posttest performance after controlling for pretest performance through a regression analysis. Second, posttest performance was predicted through performance on the scaffolding types and the contribution of feedback.

RESULTS

Psychometric Quality

Reliability

In **Table 1**, the 30 *global understanding* items from the pretest together show good reliability (α = 0.82). The same items from the posttest showed even better reliability in the second experimental and control conditions (both α = 0.89). However, in the first experimental condition, these items showed very low reliability (α = 0.36). An overview of the missing percentage values per item on the posttest are shown in **Figure 5**. For every subtest, the missing percentage values in both experimental conditions increased considerably as the test continued, indicating that the test was excessively long in these conditions; a large proportion of the children were not able to finish the subtests.

When the observations for the last four items of every subtest were excluded from the analyses, Cronbach's alpha for the *global understanding* scale exceeded 0.70 in the first experimental condition (α = 0.73) and decreased only slightly in the second experimental and control conditions (α = 0.85 and α = 0.83). Thus, in the first experimental condition, the observations made for the last few items of every subtest showed a negative effect on the reliability of the total scale. Therefore, we chose to proceed with all analyses with only the items corresponding to the six texts presented at the beginning of every post-subtest, leaving a total of 18 texts. As presented in **Table 2**, the corresponding 54 *word-form knowledge* items together showed good reliability in both experimental conditions (α = 0.90 and α = 0.91) as well as the 36 *word-meaning knowledge* items (α = 0.79 and α = 0.88). The 18 *local cohesion inference* items, which were only administered in the first experimental condition, together showed poor reliability (α = 0.47). Therefore, we cannot make any statements about the children's ability to make local cohesion inferences.

Validity

The multitrait-multimethod matrix (MTMM) for the dynamic assessment scales and NSTs is presented in **Table 3**. The scale *global understanding* shows a high correlation with the NST that measures the similar construct of text comprehension (r = 0.51) as well as the NST measuring construct vocabulary (r = 0.50). The scale *word-meaning knowledge* shows a high correlation with the NST that measures the similar construct of vocabulary (r = 0.52) and a slightly higher correlation with the NST measuring the construct of text comprehension (r = 0.59). The scale *word-form knowledge* shows a high correlation with the NST that measures the similar construct of orthography (r = 0.83) and lower correlations with the NSTs measuring less related constructs. Furthermore, from the intercorrelations between the subscales, we can conclude that *word-form knowledge* discriminates better with *global understanding* and *word-meaning knowledge* (r = 0.47 and r = 0.52) than the latter do among themselves (r = 0.68).

TABLE 1 | Reliability of the scale of global understanding.

	n items	*n* persons	α	90% CI	μ rit	μ *p*
Pretest	30	169	0.82	(0.79, 0.85)	0.39	0.60
Posttest						
Condition 1	30	47	0.36	(0.11, 0.56)	0.37	0.49
Condition 2	30	48	0.89	(0.85, 0.92)	0.49	0.52
Condition 3	30	45	0.89	(0.85, 0.93)	0.49	0.65
Shortened pretest	18	169	0.74	(0.69, 0.79)	0.41	0.60
Shortened posttest						
Condition 1	18	47	0.73	(0.62, 0.81)	0.43	0.56
Condition 2	18	48	0.85	(0.80, 0.90)	0.51	0.54
Condition 3	18	45	0.83	(0.76, 0.88)	0.50	0.65

Learning Potential and Instructional Needs

To determine children's learning potential, posttest performance on *global understanding* was predicted through the conditions after controlling for pretest performance. Compared to the control condition, both experimental conditions showed a negative effect on posttest performance, indicating that scaffolding deteriorated posttest performance (see **Table 4**).

To determine whether we were still able to provide information about children's instructional needs, posttest performance on *global understanding* was predicted by performance on the scaffolding tasks and the contribution of feedback. Scaffolding was operationalized as a percentage of the items that were answered correctly during the first attempt. Feedback was operationalized as a percentage of the items that were answered incorrectly during the first attempt and correctly during the second attempt. Since both experimental conditions received word-level scaffolding and feedback, we chose to include both groups in the same model, with the condition as a control variable and the first condition as the reference category. The predictors explained a significant part of the variation in posttest performance, R^2 = 0.477, $F(5, 89)$ = 16.25, p < 0.001. From the results shown in **Table 5**, we can conclude that scaffolding on both *word-form knowledge* (β = 0.209, p = 0.073) and *word-meaning knowledge* (β = 0.784, p < 0.001) was a relevant

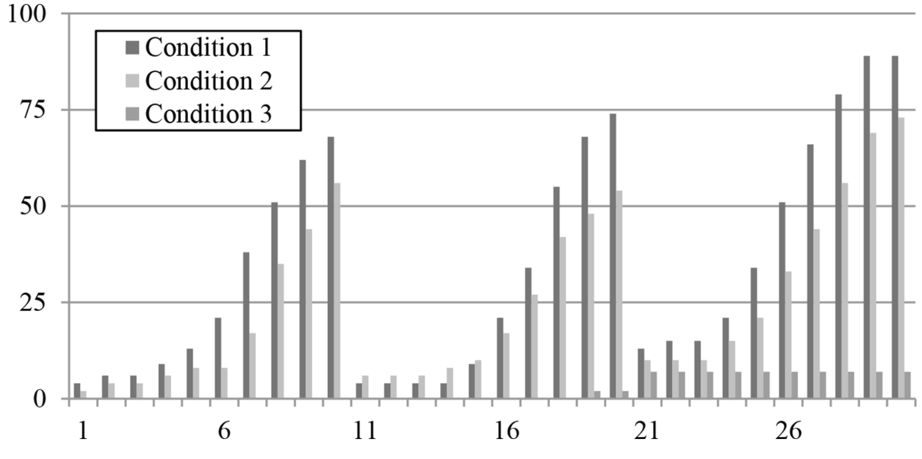


FIGURE 5 | Percentage of missing values per item from the posttest.

predictor of global understanding. The feedback showed no significant contribution. Although no significant effects could be proved, the high standardized beta for feedback on *word-meaning knowledge* suggests the potential relevance of this type of feedback (β = 0.212, p = 0.240).

Since the experimental conditions did not perform better on *global understanding* than the control condition, children's learning potential could not be assessed. Within the experimental conditions, however, scaffolding proved to be relevant for explaining performance on *global understanding*. Therefore, we were able to provide diagnostic information on children's text comprehension skills.

DISCUSSION

In order to define design principles for fine-grained assessment of text comprehension skills, a computer-based dynamic assessment based on the proposed assessment model was developed and evaluated in an experimental design. We examined to what extent we were able to measure a combination of the different aspects of the reading process by evaluating the quality of all scales. We found that a large proportion of the children in both experimental conditions were unable to finish the subtests of the posttest, indicating that these tests were excessively long. In relation to *global understanding*, the test length showed a negative effect on the reliability of the scale in the experimental condition, where children received both word- and sentence-level scaffolding and feedback. Thus, in particular, scaffolding and feedback on the sentence level (i.e., *local cohesion inferences*) resulted in inconsistent response behavior on the *global understanding* scale, indicating concentration and motivation challenges. The inclusion of six texts per subtest proved to be the maximum for obtaining a reliable *global understanding* scale. Proceeding with the analyses with only these texts, w*ord-form knowledge* and *word-meaning knowledge* were also evaluated to be reliable scales.

TABLE 2 | Reliability of the scaffolding scales.

	***n* items**	***n* persons**	**α**	**90% CI**	***μ* rit**	***μ p***
Word-form knowledge						
Condition 1	54	47	0.90	(0.86, 0.93)	0.31	0.60
Condition 2	54	48	0.92	(0.89, 0.95)	0.39	0.60
Word-meaning knowledge						
Condition 1	36	47	0.79	(0.72, 0.86)	0.30	0.73
Condition 2	36	48	0.88	(0.83, 0.92)	0.40	0.71
Local cohesion inferences						
Condition 1	18	47	0.47	(0.27, 0.64)	–	–

TABLE 4 | Regression coefficients predicting posttest performance controlled for pretest performance.

	Unstandardized coefficients		**Standardized coefficients**	
	B	**SE**	**β**	***p***
Pretest	0.637	0.073	0.589	<0.001
Condition 1*	−2.074	0.673	−0.242	0.002
Condition 2*	−2.186	0.668	−0.256	0.001

*R^2 = 0.387, F(3, 136) = 28.61, p < 0.001. *Condition 3 served as the reference category.*

TABLE 5 | Regression coefficients predicting posttest performance on global understanding.

	Unstandardized coefficients		**Standardized coefficients**	
	B	**SE**	**β**	***p***
Word-form knowledge				
Scaffolding	4.752	2.619	0.209	0.073
Feedback	3.844	4.488	0.090	0.394
Word-meaning knowledge				
Scaffolding	19.256	4.635	0.784	<0.001
Feedback	8.049	6.801	0.212	0.240
Control variables				
Condition 2*	0.094	0.602	0.012	0.877

*Scaffolding was operationalized as a percentage of the items that were answered correctly during the first attempt. Feedback was operationalized as a percentage of the items that were answered incorrectly during the first attempt and correctly during the second attempt. *Condition 1 served as the reference category.*

TABLE 3 | Multitrait-multimethod matrix for the dynamic assessment scales and nationally standardized tests.

		Dynamic assessment (DA)		
Method	**Trait**	**Global understanding**	**Word-meaning knowledge**	**Word-form knowledge**
DA	Global understanding	(0.88)		
	Word-meaning knowledge	0.68	(0.91)	
	Word-form knowledge	0.47	0.52	(0.93)
NST	Text comprehension	**0.51**	0.59	0.50
	Vocabulary	0.50	**0.52**	0.46
	Orthography	0.34	0.46	**0.81**
	Mathematics	0.33	0.31	0.32

The dynamic assessment scales are based on the posttest data for Conditions 1 and 2 together. The values in parentheses represent the reliability (α) of the scale. The values of the validity diagonal are shown in bold and are expected to be high. All correlations were found to be statistically significant different from zero (p < 0.05).

The *local cohesion inferences* scale was found to be highly unreliable on the computer-based dynamic assessment, though it was found to be perfectly reliable and well-constructed when administered in isolation with a paper-based test in the preliminary research. Two possible explanations are conceivable for this difference. First, these items could function differently on a computer-based test than on a paper-based test. To find the correct response to the items, it is necessary to read the text and find the relevant text phrases. Reading a text presented on a computer screen usually entails higher cognitive workload than reading a text presented on paper (Mangen et al., 2013). Second, the items could function differently when administered in isolation, compared to when they are administered together in a series of tasks. For every text, the children were first presented with the items regarding their word knowledge. In order to find the correct responses to these items, reading the text could have helped, though it was not necessary. Finding the correct response to the *local cohesion inferences* item, however, required the children to use information from the text. Moreover, making inferences is perceived as a higher-level ability and is, therefore, more cognitively demanding than activating word knowledge, which is perceived as a lower-level ability. The change in the required approach to problem-solving might have caused confusion or motivational problems.

We, therefore, concluded that the underlying constructs measured by the scales *global understanding* and *word-meaning knowledge* overlapped considerably. Also, both scales showed almost equal coherence with other tests that measured text comprehension and vocabulary in isolation. This could be explained by the essential role of vocabulary in text comprehension as well as by the inability of the other tests to measure text comprehension and vocabulary as separate abilities, since these abilities continuously interact and influence each other (Verhoeven et al., 2011; Oakhill and Cain, 2012). Correlations of 0.80 between tests for reading comprehension and vocabulary are not uncommon (e.g., Tomesen et al., 2017, 2018), and this supports the assumption that different reading abilities should be measured in conjunction.

To determine whether we were able to measure learning potential, the posttest performance on *global understanding* was compared for the experimental conditions versus the control condition after controlling for the pretest performance. On average, those children who received scaffolding and feedback were found to perform slightly worse than those who received no scaffolding or feedback. This could be variously explained.

As discussed earlier, children might benefit differently from instruction. Therefore, linking learning potential to specific characteristics might provide more meaningful information, since it allows for the identification of groups with similar educational needs. However, the sample size of the present study was too small to determine learning potential for smaller groups. A larger sample size would also allow for the estimation of test scores with the use of item-response theory models. These models can provide more accurate scores, as they take into account the difference between the difficulty of an item and the ability level of a student.

The most likely explanation for the lack of finding a positive effect of scaffolding and feedback concerns the possible incomparability between pre- and posttest performance on *global understanding* as well as between the experimental and control groups. When the *global understanding* task is integrated in a series of tasks, the conditions under which the children perform change. The required change in approach to problem solving between the different tasks, as pointed out earlier with respect to the *local cohesion inferences* scale, can affect children's performance or the difficulty of the tasks. Shifting the focus from measuring learning potential to measuring instructional needs would provide teachers with more valuable information.

Another probable explanation is that the information retrieved from the scaffolding and feedback was not used for the *global understanding* task. The children did not receive explicit information about the structure of the test; consequently, they themselves had to realize that they could use the previously collected information to solve the task. Moreover, previous research suggests that computer-delivered elaborated feedback is likely to be neglected in a low-stakes assessment setting on higher-order processes of text comprehension (Golke et al., 2015). Motivating children to use the information provided might address this problem.

To determine whether we were able to provide information on children's instructional needs, performance on *global understanding* was predicted by performance on scaffolding and the contribution of feedback. Scaffolding on *word-form knowledge* and *word-meaning knowledge* proved to be relevant for *global understanding*. The contribution of feedback on *word-meaning knowledge* could not be proved, although there were indications that it might be proved in a larger sample. Previous research has indicated the efficacy of using pictures when learning new words (Gruhn et al., 2019, under review). Therefore, further research is necessary to determine the contribution of this type of feedback within the assessment model. Feedback on *word-form knowledge* showed no contribution to the prediction of *global understanding*. This might be due to the lack of repetition (Gruhn et al., 2019).

Thus, some important information on children's instructional needs could be provided. However, further research is required, since inference-making skills could not be reliably assessed, and being able to integrate multiple sentences is essential for achieving *global understanding* of a text (Best et al., 2005).

CONCLUSION

Based on the findings we can conclude that the assessment model can be used as framework for fine-grained assessment of text comprehension skills when some design principles are taken into account. First, children should be informed about the test structure in advance. In this way, they can be explicitly instructed to use information retrieved from the scaffolding or feedback. Second, (sub)tests should be relatively short so as to avoid fatigue effects that cause biased results. We recommend a maximum of six short (ca. 200 words) texts per subtest. Third, a pretest where one's ability is measured in isolation might not be a good baseline for establishing learning potential. Further research is required, however, since the inability to establish learning potential in the present study might have had other causes. In any case, comparability of response behavior elicited by pre- and posttest measures should be examined beforehand. Fourth, the negative effects of changes in the required approach to problem-solving or the cognitive workload between different tasks should be diminished. These effects might be reduced by presenting a visual indication of the level of difficulty or a sign reflecting the task type of every task. In the present study, the change from the *word-meaning knowledge* task to the *local cohesion inference* task seems to cause problems in particular. In addition to a visual indicator or (warning) sign, reducing the cognitive workload for this specific task may also contribute. This could be achieved by adding a new, in-between task that serves as an extra intermediate step or by directly highlighting the relevant text passages instead of only after an incorrect response. However, attention must be paid to the influence of such adjustments on the validity of the task itself and the following tasks.

To conclude, we have tried to bridge the first gap between theory and assessment by evaluating the theoretically based assessment model for fine-grained assessment of text comprehension skills in practice. We were able to measure a combination of different aspects of the reading process. Furthermore, we suggested that it might be more valuable to focus on instructional needs rather than on learning potential. Through the design principles discussed, we can move further toward fine-grained assessment of text comprehension skills.

ETHICS STATEMENT

We obtained written informed consent from all participating schools and the children's parents were informed about the study by letter. The parents had the opportunity to refuse participation of their child in the study. Active written informed parental consent was not considered necessary because we only collected non-special personal data; biographical information or any other private or health-related data were not part of our study. This particular sub-study was conducted under a research grant of the Netherlands Initiative for Education Research (NWO/NRO dossiernummer 405-15-548), which was approved by a scientific and ethical review committee. Therefore, because of institutional (Cito) and national (NOW/NRO) guidelines and regulations, the separate approval by an Ethics Committee was not required.

AUTHOR'S NOTE

The following data were collected in our study: (1) Correct/incorrect scores on a dynamic assessment for text comprehension and (2) test scores for reading comprehension, spelling, vocabulary and math from a Dutch Student Monitoring System.

AUTHOR CONTRIBUTIONS

MO is a PhD student. JK is the daily supervisor. TE is the supervisor of the project. The authors are equally responsible for the content.

FUNDING

Our study was conducted under a research grant of the Netherlands Initiative for Education Research (NWO/NRO filing reference 405-15-548) and was approved by a scientific review committee.

REFERENCES

Best, R. M., Rowe, M., Ozuru, Y., and McNamara, D. S. (2005). Deep-level comprehension of science texts: the role of the reader and the text. *Top. Lang. Disord.* 25, 65–83. doi: 10.1097/00011363-200501000-00007

Brown, A., and Ferrara, R. (1985). "Diagnosing zones of proximal development" in *Culture, communication, and cognition: Vygotskian perspectives.* ed. J. Wertsch (Cambridge: Cambridge University Press), 273–305.

Caffrey, E., Fuchs, D., and Fuchs, L. S. (2008). The predictive validity of dynamic assessment: a review. *J. Spec. Educ.* 41, 254–270. doi: 10.1177/0022466907310366

Cain, K., Compton, D. L., and Parrila R. K. (eds.) (2017). *Theories of reading development,* Vol. 15. (Amsterdam: John Benjamins).

Cain, K., and Oakhill, J. (2006). Profiles of children with specific reading comprehension difficulties. *Br. J. Educ. Psychol.* 76, 683–696. doi: 10.1348/000709905X67610

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56, 81–105.

Campione, J. C., and Brown, A. (1987). "Linking dynamic assessment with school achievement" in *Dynamic assessment: An interactional approach to evaluating learning potential.* ed. C. S. Lidz (New York: Guilford), 82–115.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* 2nd edn. (Hillsdale, NJ: Lawrence Erlbaum Associates).

Colenbrander, D., Kohnen, S., Smith-Lock, K., and Nickels, L. (2016). Individual differences in the vocabulary skills of children with poor reading comprehension. *Learn. Individ. Differ.* 50, 210–220. doi: 10.1016/j.lindif.2016.07.021

Daugaard, H. T., Cain, K., and Elbro, C. (2017). From words to text: inference making mediates the role of vocabulary in children's reading comprehension. *Read. Writ.* 30, 1773–1788. doi: 10.1007/s11145-017-9752-2

Dennen, V. P., and Burner, K. J. (2008). "The cognitive apprenticeship model in educational practice" in *Handbook of research on educational communications and technology.* 3rd edn. eds. J. M. Spector, M. D. Merrill, J. van Merrienböer, and M. P. Driscoll (New York, NY: Lawrence Erlbaum Associated), 425–439.

Dörfler, T., Golke, S., and Artelt, C. (2009). Dynamic assessment and its potential for the assessment of reading competence. *Stud. Educ. Eval.* 35, 77–82. doi: 10.1016/j.stueduc.2009.10.005

Dörfler, T., Golke, S., and Artelt, C. (2017). "Evaluating prerequisites for the development of a dynamic test of reading competence: feedback effects on reading comprehension in children" in *Competence assessment in education.* eds. D. Leutner, J. Fleischer, J. Grünkorn, and E. Klieme (Cham: Springer), 487–503.

Elliott, J. (2003). Dynamic assessment in educational settings: realising potential. *Educ. Rev.* 55, 15–32. doi: 10.1080/00131910303253

Evers, A., Sijtsma, K., Lucassen, W., and Meijer, R. R. (2010). The Dutch review process for evaluating the quality of psychological tests: history, procedure, and results. *Int. J. Test.* 10, 295–317. doi: 10.1080/15305058.2010.518325

Fani, T., and Rashtchi, M. (2015). Dynamic assessment of reading comprehension ability: group or individualized. *Educ. J.* 4, 325–331. doi: 10.11648/j.edu.20150406.11

Feng, M. (2009). Towards assessing students' fine grained knowledge: using an intelligent tutor for assessment. (Doctoral dissertation, Carnegie Mellon University).

Fuchs, D., Fuchs, L. S., and Compton, D. L. (2012). Smart RTI: a next-generation approach to multilevel prevention. *Except. Child.* 78, 263–279. doi: 10.1177/001440291207800301

Golke, S., Dörfler, T., and Artelt, C. (2015). The impact of elaborated feedback on text comprehension within a computer-based assessment. *Learn. Instr.* 39, 123–136. doi: 10.1016/j.learninstruc.2015.05.009

Gruhn, S., Segers, E., and Verhoeven, L. (2019). The efficiency of briefly presenting word forms in a computerized repeated spelling training. *Read. Writ. Q.* 35, 225–242. doi: 10.1080/10573569.2018.1526725

Keenan, J. M., Betjemann, R. S., and Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: differential dependence on decoding and oral comprehension. *Sci. Stud. Read.* 12, 281–300. doi: 10.1080/10888430802132279

Kintsch, W. (1988). The use of knowledge in discourse processing: a construction-integration model. *Psychol. Rev.* 95, 163–182.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* (Cambridge, UK: Cambridge University Press).

Kleinsz, N., Potockib, A., Ecallea, J., and Magnana, A. (2017). Profiles of French poor readers: underlying difficulties and effects of computerized training programs. *Learn. Individ. Differ.* 57, 45–57. doi: 10.1016/j.lindif.2017.05.009

Lantolf, J. P., and Poehner, M. E. (2007). "Dynamic assessment" in *Encyclopedia of language and education: Language testing and assessment.* eds. E. Shohamy, and N. H. Hornberger, Vol. 7. (Berlin: Springer), 273–285.

Lidz, C. S., and Elliott, J. G. (2000). *Dynamic assessment: Prevailing models and applications.* (Amsterdam: Elsevier).

Magno, C. (2010). The effect of scaffolding on children's reading speed, reading anxiety, and reading proficiency. *TESOL J.* 3, 92–98. Retrieved at: https://www.tesol-international-journal.com/wp-content/uploads/2013/11/A6_V3_TESOL.pdf

Mangen, A., Walgermo, B. R., and Brønnick, K. (2013). Reading linear texts on paper versus computer screen: effects on reading comprehension. *Int. J. Educ. Res.* 58, 61–68. doi: 10.1016/j.ijer.2012.12.002

Nation, K., and Snowling, M. (1997). Assessing reading difficulties: the validity and utility of current measures of reading skill. *Br. J. Educ. Psychol.* 67, 359–370.

Oakhill, J. V., and Cain, K. (2012). The precursors of reading ability in young readers: evidence from a four-year longitudinal study. *Sci. Stud. Read.* 16, 91–121. doi: 10.1080/10888438.2010.529219

Pander Maat, H., Kraf, R., Van den Bosch, A., Dekker, N., Van Gompel, M., Kleijn, S., et al. (2014). T-Scan: a new tool for analyzing Dutch text. *Computat. Linguist. Neth. J.* 4, 53–74.

Perfetti, C. A. (1999). "Comprehending written language: a blueprint of the reader" in *The neurocognition of language.* eds. C. M. Brown, and P. Hagoort (Oxford, UK: Oxford University Press), 167–208.

Perfetti, C. A. (2017). "Lexical quality revisited" in *Developmental perspectives in written language and literacy: In honor of Ludo Verhoeven.* eds. E. Segers, and P. van den Broek (Amsterdam: John Benjamins), 51–68.

Perfetti, C. A., and Hart, L. (2001). "The lexical basis of comprehension skill" in *On the consequences of meaning selection.* ed. D. Gorfien (Washington, DC: American Psychological Association), 67–86.

Perfetti, C. A., Landi, N., and Oakhill, J. (2005). "The acquisition of reading comprehension skill" in *The science of reading: A handbook.* eds. M. J. Snowling, and C. Hulme (Oxford: Blackwell Publishing), 227–247.

Poehner, M. E., and Lantolf, J. P. (2013). Bringing the ZPD into the equation: Capturing L2 development during computerized dynamic assessment. *Lang. Teach. Res.* 17, 323–342. doi: 10.1177/1362168813482935

Segers, E., and Verhoeven, L. (2016). How logical reasoning mediates the relation between lexical quality and reading comprehension. *Read. Writ.* 29, 577–590. doi: 10.1007/s11145-015-9613-9

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Read. Res. Q.* 16, 32–71.

Sternberg, R. J., and Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential.* (Cambridge: Cambridge University Press).

Swart, N. M., Muijselaar, M. M. L., Steenbeek-Planting, E. G., Droop, M., De Jong, P. F., and Verhoeven, L. (2017). Cognitive precursors of the developmental relation between lexical quality and reading comprehension in the intermediate elementary grades. *Learn. Individ. Differ.* 59, 43–54. doi: 10.1016/j.lindif.2017.08.009

The Netherlands Institute for Social Research (2016). Statusscores. Retrieved from: May 28, 2018, https://www.scp.nl/Onderzoek/Lopend_onderzoek/A_Z_alle_lopende_ onderzoeken/Statusscores

Tomesen, M., Engelen, R., and Hiddink, L. (2018). *Wetenschappelijke verantwoording Begrijpend Lezen 3.0 voor groep 7.* (Arnhem: Cito).

Tomesen, M., Weekers, A., Hiddink, L., and Jolink, A. (2017). *Wetenschappelijke verantwoording Begrijpend Lezen 3.0 voor groep 6.* (Arnhem: Cito).

Tzuriel, D. (2000). Dynamic assessment of young children: educational and intervention perspectives. *Educ. Psychol. Rev.* 12, 385–435. doi: 10.1023/A:1009032414088

Van der Kleij, F. M., Feskens, R. C., and Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: a meta-analysis. *Rev. Educ. Res.* 85, 475–511. doi: 10.3102/0034654314564881

Van Dijk, T. A., and Kintsch, W. (1983). *Strategies of discourse comprehension.* (New York: Academic Press).

Verhoeven, L., and Perfetti, C. (2008). Advances in text comprehension: model, process and development. *Appl. Cogn. Psychol.* 22, 293–301. doi: 10.1002/acp.1417

Verhoeven, L., Van Leeuwe, J., and Vermeer, A. (2011). Vocabulary growth and reading development across the elementary school years. *Sci. Stud. Read.* 15, 8–25. doi: 10.1080/10888438.2011.536125

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* (Cambridge, MA: Harvard University Press).

Conflict of Interest Statement: TE and JK were employed by Cito in The Netherlands. They are affiliated with Stichting Cito (Foundation Cito), the not-for-profit part of Cito that is dedicated to applied scientific research into educational measurement. MO was employed at the University of Twente.

frontiers in Psychology

CONCEPTUAL ANALYSIS
published: 03 July 2019
doi: 10.3389/fpsyg.2019.01522

Online Diagnostic Assessment in Support of Personalized Teaching and Learning: The eDia System

Benő Csapó[1]* *and Gyöngyvér Molnár*[2]

[1] *MTA-SZTE Research Group on the Development of Competencies, University of Szeged, Szeged, Hungary,*
[2] *Department of Learning and Instruction, University of Szeged, Szeged, Hungary*

The aims of this paper are: to provide a comprehensive introduction to eDia, an online diagnostic assessment system; to show how the use of technology can contribute to solve certain crucial problems in education by supporting the personalization of learning; and to offer a general reference for further eDia-based studies. The primary function for which the system is designed is to provide regular diagnostic feedback in three main domains of education, reading, mathematics, and science, from the beginning of schooling to the end of the 6 years of primary education. The cognitive foundations of the system, the assessment frameworks, are based on a three-dimensional approach in each domain, distinguishing the psychological (reasoning), the application, and the disciplinary (curricular content) dimensions of learning. The frameworks have been carefully mapped into item banks containing over a 1,000 innovative (multimedia-supported) items in each dimension. The online assessments were piloted, and the system has been operating in experimental mode in over 1,000 schools for several years. This paper outlines the theoretical foundations of the eDia system and summarizes how results from research on the cognitive sciences, learning and instruction, and technology-based assessment have been integrated into a working system designed to assess a large population of students. The paper describes the main functions of eDia and discusses how it supports item writing, constructing tests, online test delivery, automated scoring, data processing, scaling and the provision of feedback both for students and teachers. It shows how diagnostic assessments can be implemented in school practice to facilitate differentiated instruction through regular measurements and to provide instruments for teachers to make formative assessments. Beyond its main function (supporting development toward personalizing education), the eDia platform has been used for assessments in a number of areas from pre-school to higher education both in Hungary and in a number of other countries as well. The paper also reviews results from eDia-based studies and highlights how technology-based assessment extends the possibilities of educational research by making more constructs measurable.

Keywords: technology-based assessment, online assessment, diagnostic assessment, assessment framework, item banking

OPEN ACCESS

Edited by:
Frank Goldhammer,
German Institute for International Educational Research (LG), Germany

Reviewed by:
Birgit Schütze,
University of Münster, Germany
Diego Zapata-Rivera,
Educational Testing Service, United States

****Correspondence:***
Benő Csapó
csapo@edpsy.u-szeged.hu

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *17 December 2018*
Accepted: *17 June 2019*
Published: *03 July 2019*

Citation:
Csapó B and Molnár G (2019) Online Diagnostic Assessment in Support of Personalized Teaching and Learning: The eDia System. Front. Psychol. 10:1522. doi: 10.3389/fpsyg.2019.01522

INTRODUCTION

The eDia online assessment system has been built and developed by the Centre for Research on Learning and Instruction, University of Szeged. The principal function for which the system is designed is to provide regular diagnostic information in three main domains of education, reading, mathematics, and science, from the beginning of schooling to the end of the 6 years of primary education. In its present form, the eDia system is an integrated assessment system that is based on sophisticated frameworks and supports assessment processes from item development through test administration and data analyses to well-interpretable feedback. It is one realization of the "integrated, learning-centered assessment systems" envisioned by Pellegrino and Quellmalz (2010).

One of the main challenges of school education stems from the fact that students are different. Looking at the problem from a historical perspective, two main approaches may be identified as school systems have attempted to respond to this challenge: (1) selecting students (ability grouping, tracking, etc.) in the hope that homogeneous classrooms can be set up and (2) accepting different students for heterogeneous classrooms, then differentiating instruction to adjust teaching to the different individual needs of the students (personalization, individualization, etc.). The first option has failed, mostly for two reasons: (1) students are different not only in one dimension but also in a number of different ways, with the differences changing dynamically over time; therefore, (2) the intention of selection has generally resulted in social selection (segregation) with numerous negative side effects. The second option is more promising, and a number of progressive initiatives have emerged in recent decades. However, there have also been a great many difficulties that have stood in the way of personalizing learning; among these, the most prominent is continuously identifying the critical differences between students, differences that determine successful learning options. The most crucial issue in teaching a heterogeneous classroom is teaching students with temporary or permanent difficulties in learning, thus requiring that the difficulties that block their progress be identified.

From a cognitive point of view, the core of the problem was best conceptualized by Ausubel in his frequently cited observation: "The most important single factor influencing learning is what the learner already knows. Ascertain this and teach him accordingly" (Ausubel, 1968, p. vi). As simple as this idea is, it is equally as difficult to implement in heterogeneous classrooms. To realize this in practice, teachers should know "what the learner already knows." The problem of "knowing what students know," as has been formulated by several authors (Pellegrino et al., 2001; Opfer et al., 2012), has been solved in general, but making this knowledge useable in practice, teachers should know in "real time," or at least should receive feedback with sufficient frequency to be able to adjust teaching to the knowledge currently possessed by learners. It is clear that due to material costs and human resources requirements, systematic large-scale diagnostic assessments cannot be conducted with traditional instruments.

In this paper, we first outline the theoretical foundations of the eDia system, including the role of diagnostic assessment, the content of assessment, and the ways to use feedback. Then, we introduce the eDia system, describe its structure, and highlight how technology serves its functions. Finally, we review research studies that have been carried out using eDia.

Throughout this paper, we emphasize that there are a number of innovations that technology brings into numerous aspects of instructional processes, including assessment. However, currently, there is still unexploited potential in the use of technology, including the possibilities of personalizing learning, adjusting teaching and learning processes to the individual needs of students. From a cognitive point of view, if students are always taught what they are prepared for (as Vygotsky's theory of the zone of proximal development proposes), then they will better comprehend and master the teaching material. From an affective perspective, if each student individually always faces an optimally challenging learning task (as Csíkszentmihályi's theory of optimal experiences proposes, see Csíkszentmihályi, 2000), both boredom and anxiety are eliminated from learning processes and maintains motivation. The optimal level of challenge supports students' need for competence, which has a positive impact on students' intrinsic motivation as well (Ryan and Deci, 2000a,b). We notice here that large item banks also allow personalization of assessment so that each student receives tests adjusted to their actual developmental level (adaptive testing), thus reducing anxiety in the assessment process as well. Both cognitive and affective demands require regular, personalized feedback, which is what eDia is designed for.

THEORETICAL FRAMEWORK

The eDia system constitutes the core of a complex, novel educational model which synthesizes a number of progressive initiatives to improve education. It is designed to support learning and development in the first phase of schooling and takes into account certain realities that determine the possibilities of using technologies. We consider three sets of conditions under which problems must be solved.

1. We assume that the role of teachers remains central in the teaching and learning processes. Their personal presence is needed in the classroom, especially in the first year of schooling. Therefore, the technology in the proposed model is not meant to replace the teacher, but to provide diagnostic tools to support their work. With such diagnostic tools, teachers will be empowered to improve their own work by experimenting, modifying the way they teach and assessing the impact, as research-based teacher education (Westbury et al., 2005; Munthe and Rogne, 2015) prepares them for such activities and as required by evidence-based educational practice (Slavin, 2002).
2. The second reality is the large differences between pupils. We assume, based on evidence from numerous analyses that heterogeneous, inclusive schools and classrooms are

more efficient, with both quality and equity potentially ensured simultaneously; however, teaching in heterogeneous classes may be more difficult. The major challenge is to adjust instruction to the individual needs of every student. Diagnostic assessment may help, as it provides information on the actual developmental level of each pupil.

3. We assume that regular feedback is essential for learning. A major trend to provide students with proper feedback has been promoted through formative assessments. We agree with its importance, but at the same time, we assume that teachers are not able to observe every major aspect of learning without an objective assessment instrument. Furthermore, traditional paper-based instruments are not suitable for rapid and frequent feedback. Technology-based diagnostic assessments may fill this gap.

Given these conditions, four major research trends offer results for integration and synthesis that serve as a theoretical foundation for a complex online diagnostic assessment system. (1) In research and development, there is a shift from summative to formative assessment, which provides immediate feedback and direct support for learning. (2) Technology-based assessment has shown enormous progress in the past decade, and ICT infrastructure in schools has improved so that assessment can enter into everyday school practice. (3) Progress in cognitive and educational psychology has produced results which have not yet been exploited in practice and which may contribute to a solution for certain crucial problems, especially in the first year of schooling. (4) Finally, a number of promising models for personalizing learning has had limited influence on practice, mostly because of the lack of easy-to-use assessment instruments. Although efforts within this latter (4) trend highlight the need for regular diagnostic feedback and the reformed teaching methods provide adequate educational context for the assessments, in this section, we only deal in detail with the first (1–3) trends as they have determined the development of the eDia system more directly.

Formative and Diagnostic Assessment

Large-scale international assessment programs (Trends in International Mathematics and Science Studies – TIMSS, Progress in International Reading Literacy Study – PIRLS, and Program for International Student Assessment – PISA) have had an immense impact on the development of educational systems in many different ways and have inspired the introduction or expansion of national assessment programs. These programs have also advanced testing in a number of areas, including framework development, test administration, data analyses, and reporting. This progress has also highlighted some deficiencies in educational assessment from the perspective of practice as well, for example, the long time between test administration and feedback, the limited usefulness of summative test results with regard to personalized intervention, and the lack or limitations of student-level feedback in general. Another source of dissatisfaction with testing has been the way summative tests have been used in certain countries, especially for high-stakes assessments, e.g., for test-based accountability. These types of testing have caused some negative effects, such as teaching for testing and test score inflation (see, e.g., Koretz, 2018), as well as harmful influence on school climate and teacher stress (Saeki et al., 2018).

These deficiencies have lent a new impetus for other directions in the development of educational assessment and shifted the focus of attention from summative to formative assessment (Clarke, 2001, 2005; Ainsworth and Viegut, 2006; Bennett and Gitomer, 2009; Bennett, 2011; Sheard and Chambers, 2014), or assessment for learning, as it is often called (Black et al., 2003; Hattie and Brown, 2007; Heitink et al., 2016), or diagnostic assessment, to use yet another term (Leighton and Gierl, 2007). There are many different ways formative assessment is used in practice, but a common feature of these assessments is that they reflect students' learning needs, facilitate understanding in a given context and provide students with immediate feedback (Black and Wiliam, 1998a,b; Black et al., 2004; Good, 2011). There is no sharp distinction between formative and diagnostic assessment, nor does a universal definition for diagnostic assessment exist. However, it is usually described as a kind of assessment which focuses on problems, explores possible difficulties, assesses if students are prepared for a learning task, and thus may measure prerequisite knowledge as well. Furthermore, diagnostic assessment is often followed by a kind of "therapy": compensatory instruction to eliminate obstacles and offer various forms of supportive activities (e.g., in mathematics: Brendefur et al., 2018), which facilitates data-based decision making (e.g., in reading: Filderman et al., 2018).

One typical and most traditional form of formative assessment takes place in the context of classroom interaction, with evaluation based on teachers' observation and personal judgment. Further forms are evaluations of students' work and learning artifacts (performances, presentations, essays, worksheets, projects, documents, lab results, etc.). Although there is a need for frequent personal feedback from teachers, the subjective nature has prompted the use of objective instruments; thus, formative tests have been proposed for this purpose. As these tests have been customized and adjusted to contexts and actual needs, they have usually been teacher-made tests of questionable psychometric quality. Formative tests have been used most systematically in personalized models of instruction, but in any case, their production, administration, and scoring have required immense resources. The use of technology has been proposed to solve these problems, to support certain aspects of the assessments (Feng and Heffernan, 2005; Brown et al., 2008; Feng et al., 2009) or to devise comprehensive assessment systems (Perie et al., 2009).

Evolution of Technology-Based Assessment

Although technology-based assessment (TBA) is almost as old as the computer itself, modern TBA has a much shorter history. Its potential in assessment has been clear for decades, but it has required several initiatives and the development of the infrastructure at schools to fulfill its promise. We review here only a few major projects and programs that have aided in the realization of eDia as well.

The European Union has launched several initiatives to modernize education, including the expansion of educational assessments to new areas with new technologies. The EU's Joint Research Centre has organized conferences and workshops to collect experience with TBA projects (Scheuermann and Guimarães Pereira, 2008). One such workshop was held in Reykjavik, Iceland, in September–October 2008 with the participation of over 100 experts presenting several parallel developments (Scheuermann and Björnsson, 2009). Among other software, the TAO program (open source software developed by the Centre de Recherche Public Henri Tudor and EMACS, University of Luxembourg) was introduced in several presentations, indicating that it was not only being used in the PISA studies but also in national initiatives as well (Csapó et al., 2009; Haldane, 2009). The MicroDYN approach (Greiff and Funke, 2009), which later became the core of the PISA 2012 problem-solving assessment and which is also implemented in eDia, was also presented at this meeting. In a volume based on the workshop presentations, three chapters summarized the results of the PISA Computer-Based Assessment of Science by authors from the participating countries (Iceland, Korea, and Denmark; see Halldórsson et al., 2009; Lee, 2009; Sørensen and Andersen, 2009). A chapter in the same volume by Kozma (2009) was also published, which was a call for action to assess and teach the 21st-century skills, a manifesto of the program started around that time.

The Assessment and Teaching of 21st-Century Skills (ATC21S) project was located at the intersection of two major trends in research and development: the need to re-define the purpose of education in the new millennium with a greater focus on the skills required in modern societies and to make these skills measurable through TBA. In the first phase of the project, four working groups were formed to define the targeted skills (Binkley et al., 2012) and to explore methodological, psychometric (Wilson et al., 2012), and technological (Csapó et al., 2012) issues, as well as contextual and environmental issues (Scardamalia et al., 2012). The volume that published the results contained a further chapter on the policy frameworks for the assessments (Darling-Hammond, 2012). In the second phase, the project focused on two prominent and closely related 21st-century skills, collaborative problem-solving and learning in digital networks (Griffin and Care, 2015), thus also contributing to the theoretical and empirical foundations for the 2015 PISA collaborative problem-solving assessment.

The PISA assessments have had an impact on the development of TBA in two major ways: (1) they have advanced the technological background and (2) they have tested the preparedness of individual countries for the assessments, identified deficiencies and exercised some pressure to ensure the necessary conditions to make large-scale TBA possible. The application of TBA started in 2006, when Computer-Based Assessment of Science was an optional domain (OECD, 2010). Only three countries completed the assessments (Denmark, Iceland, and Korea), but this provided an impetus for TBA within PISA. In 2009, the assessment of digital reading was an optional domain. Altogether countries participated, making the comparison of achievement in print and digital reading possible and exploring the new information-processing demands of networking and hyperlinking (OECD, 2011).

The 2012 PISA cycle brought a breakthrough in two respects. First, although paper-based tests remained the main delivery method, the TBA version of assessments was offered as an option for reading and mathematics, making the two delivery methods comparable and linking paper-based and TBA achievement (OECD, 2013). Second, in this cycle, dynamic (creative) problem-solving was the fourth, innovative assessment domain; it used simulation and interaction for the first time on PISA (OECD, 2014). This assessment has had a further impact on the development of TBA. The members of the problem-solving expert group continued meeting, invited further researchers in the field, and published an edited volume, which reported a number of further applications of and innovation in TBA (Csapó and Funke, 2017). The computerized solutions devised for the interaction in the assessment of dynamic problem-solving were adapted and further developed; they were used in 2015 for interactive science items (OECD, 2016) and for collaborative problem-solving (OECD, 2017). In 2015, the transition of PISA to TBA was complete, with all the assessments administered by computer.

The projects and programs reviewed here have influenced the development of the eDia system in several ways. PISA re-defined the content to be measured, while ATC21S linked the skills and technology used for assessment and highlighted the importance of framework development. The technology was developed in interaction with the communities running the projects under review; the major forum, beyond several meetings at conferences, was the Szeged Workshop on Educational Evaluation, held annually at the University of Szeged between 2009 and 2016. The programs reviewed here focused on summative testing among older age groups (secondary schools), underscoring the lack of formative assessment and neglecting the needs of younger students, while recent research in education has emphasized both aspects. The experiences gained from the technological realization of these programs (e.g., the item-builder technology) have been transferred to diagnostic assessments, and eDia has extended them with a number of novel solutions (e.g., item banking, a feedback system, visualization, etc.).

Beyond the developments reviewed here, a parallel evolution took place related to computer-aided instruction (Chauhan, 2017) and intelligent tutoring systems (Kulik and Fletcher, 2016) with significant assessment and feedback components (Conejo et al., 2004). The rapid development of online learning has also advanced TBA, including progress in adaptive testing (e.g., Conejo et al., 2004) and most recently in learning analytics (Avella et al., 2016), which broadens the possibilities of assessing students' learning and forms of feedback. Strategies based on several forms of computer-aided instruction and online learning designed for older students limit the role of teachers and teach students in specific domains (see, e.g., Chi et al., 2010). They open a different route for personalization and only partially overlap with the type of assessment-based differentiation for which the eDia system is devised (as for these differences, see also Scandura, 2017).

Determining What to Measure: Three-Dimensional Frameworks for Diagnostic Assessments

Previous assessment projects have stressed the importance of defining the content of assessments, and this is even more significant for diagnostic assessments in the early phases of schooling. Diagnosis requires not only a better understanding of the teaching and learning processes but also the cognitive and affective development of pupils as well. Therefore, framework development has been a prominent component in establishing the eDia system. With a brief description of framework development, we demonstrate that only the use of technology (large item banks and assessments tailored to students' individual needs) has made it a realistic goal to differentiate the special aspects of learning by defining the three dimensions of assessments.

The reading, mathematics, and science frameworks have been based on a three-dimensional model of learning outcomes. This model takes into account the traditions of defining learning objectives (e.g., creating taxonomies, developing curricula and setting standards; see Csapó, 2004, 2010) and recent research findings in fields ranging from cognitive neuroscience (e.g., Ansari and Coch, 2006) through early childhood education (e.g., McLachlan et al., 2018) to research on teaching and learning in the domains assessed.

The most traditional dimension of learning outcomes is mastering the learning material, i.e., subject matter knowledge, represented in textbooks and defined more generally in the school curricula. This type of knowledge is the easiest for teachers to observe. The most frequently assessed and graded dimension, it is termed the *disciplinary dimension* in the diagnostic frameworks. It has been the central part of many curriculum- or textbook-oriented summative assessments as well as of the first international assessment programs. The PISA frameworks have re-defined the conception of valid knowledge and expanded the interpretation of literacy in a parallel form for the three assessment domains (e.g., OECD, 1999, 2003). The same type of knowledge is assessed in the eDia diagnostic system, which is called the *application dimension*. The third dimension focuses on students' cognitive development, the processes underlying learning, which is called the *psychological dimension* (for the cognitive foundations, see also the CBAL approach, Bennett, 2010). Although PISA also assesses disciplinary knowledge in mathematics and science, it does so through the applications, while the psychological dimension appears in the innovative domain (e.g., complex problem-solving in 2003, creative problem-solving in 2012, and collaborative problem-solving in 2015). The predecessors to TIMSS focused on knowledge defined in the curricula of the participating countries, so the main resource was disciplinary knowledge, while recent frameworks deal with content, application, and reasoning as well (see, e.g., Mullis et al., 2001, 2005) somewhat similar to the eDia framework. None of the large-scale international assessment programs can measure how well disciplinary knowledge defined in the actual curricula is mastered, but it is defined and assessed in the disciplinary dimension of the diagnostic system.

The three-dimensional frameworks for reading (Csapó and Csépe, 2012), mathematics (Csapó and Szendrei, 2011), and science (Csapó and Szabó, 2012) have been developed by experts in the particular domains and dimensions. In the three domains, a total of nine dimensions are distinguished and defined; the theoretical foundation and previous research on each one are presented in a chapter in the framework volumes. There are similarities between mathematics and science, while reading is somewhat different. The theoretical chapters are followed by the detailed frameworks developed for primary school Grades 1–6. The descriptions are illustrated by sample items showing possible computerized, multimedia-supported item formats to assess a particular dimension. These frameworks served as training materials for the item writers, who then carefully mapped the frameworks into assessment items (over 1,500 items per dimension). They were also used to familiarize the teachers who use eDia with the content of the assessment. These items were empirically piloted, and a further set of books was published, one volume for each domain with detailed descriptions of the assessment dimensions and illustrated by a larger number of items taken from the item banks in the eDia system (Csapó et al., 2015a,b,c). These books help prepare teachers to use the system, to interpret the feedback provided by eDia, and to determine the intervention concluded from the assessment results. Sample items presented in these books also demonstrate that assessing certain aspects of learning (especially the psychological dimension) would be difficult (and almost impossible in school practice) without the use of technology.

The validity of the three-dimensional model has already been empirically tested. Based on the data collected *via* the eDia system, confirmatory factor analyses were performed separately in each grade for each domain. The results confirmed that, although there are usually significant correlations between the dimensions, they assess different psychological constructs (Molnár and Csapó, submitted). The psychometric indicators for the assessments (e.g., reliability) are constantly monitored, items with poor parameters are modified or deleted from the system, and new items are added to improve coverage of the content defined in the frameworks. (Results from quality improvement processes will be published elsewhere.)

THE eDia SYSTEM

The eDia system began being built in April 2007, when researchers at the University of Szeged implemented the TAO open source software (Plichart et al., 2004) on university servers and began to explore possibilities for it in close cooperation with and with the continuous support of the developers of TAO at the Centre de Recherche Public Henri Tudor, University of Luxembourg. Several pilot studies were completed with TAO, as well as a media effect study to compare the paper-and-pencil and online administration of an inductive reasoning test (Csapó et al., 2009). Although the first results were promising, and by that time several TAO modules had been used in the PISA assessments as well, it soon became obvious that TAO had not been designed for the type of diagnostic assessment system the researchers had aimed to build. This led to a decision to develop new software from scratch optimized for the complex requirements of the diagnostic assessments.

The eDia online diagnostic assessment system can be divided into two main parts. One is the hardware infrastructure (a server farm) and the software that operates the system. This has been developed and optimized for diagnostic assessment, e.g., being continuously accessible for the entire Grade 1–6 student population (up to 600,000 students), and for the management of large item banks (with tens of thousands of items). In addition, this infrastructure can also be used for several other assessment purposes. The other part is the main content of the system, the item banks prepared for the diagnostic assessment of reading, mathematics, and science.

The eDia system is functionally ready for the implementation of systematic assessments and has operated in experimental mode since 2015. At present, there are more than 1,000 partner schools (approx. one-third of the primary schools in Hungary), where it is used on a regular basis. It contains over 25,000 items. The software has been continuously developed, with both the number of partner schools and the number of items available in the system growing.

Currently, three different testing procedures are run with eDia. There are central assessments initiated by the assessment center three times in a school year, at the beginning, in the middle, and at the end of the year. These assessments provide data to establish item parameters and normative reference points. There are teacher-initiated assessments which are used for frequent diagnostic assessments adjusted to the needs of a class or of individual students. The teachers may compile tests out of the items available in the item banks for their own assessment activities. Furthermore, there is testing for research in numerous projects using either items from the item banks or specific tests developed for research purposes.

Structure of the System: Functions to Serve the Needs of Educational Practice

Item Writing

The system contains an item builder module that makes the task of item writing as easy as writing multimedia documents. Item developers receive extensive training in the content of the assessment and in test theory and psychometrics, enabling them to master the use of the item builder module easily (Molnár et al., 2015a,b, 2018). Items are written online, with the draft versions of items undergoing several phases of review (content, language, technical fitness, and format) before they are entered into the item pool for empirical testing. A number of tools are available to support item writing, including templates and scoring schemes. Several items can be created for one stimulus or a set of closely related stimuli; these items together form the tasks. The items in a task can be moved (e.g., added to a test) together.

Test Editing

In the present mode of operating the system, tests consisting of a number of tasks form the units of the assessment. Tests may be constructed out of the tasks in several ways. Typically, booklets are formed out of the tasks, and then they can be combined variously into tests, for example, to eliminate the position effect or to optimize linking/anchoring options. Tests can be constructed with adaptive testing techniques, i.e., based on the answers given to all previous items or to items present in the last cluster, to minimize the difference between the students' ability level and the test difficulty level.

Online Test Delivery

Students complete the diagnostic tests as part of their school activity using the available school infrastructure. The tests can be done practically from any device equipped with an internet browser, but the items are optimized for keyboard, mouse, and a large screen. For central assessments, there is an approx. two-week window when eDia is open for the actual assessment. Teacher-initiated testing can take place any time teachers find it useful (at this phase, they are not influenced on how frequently they use it). Students have a specific secret assessment identification code to log into the system.

Automated Scoring

The eDia system is designed for both automated and human scoring. However, the items in the item banks that are prepared for the regular diagnostic assessments are scored automatically, with human scoring reserved for research and specific applications. Automatic scoring makes it possible to provide immediate feedback, and it is necessary for the rapid scoring of a large number of assessments. The system offers a variety of scoring options, adjusted to item type and form of response capture.

Built-In Data Processing and Statistical Analyses

The eDia system contains a statistical analytics module, which can perform every computation required by the assessment from descriptive statistics through classical test theory to IRT modeling. The computations are programmed using the open source "R" programming language and are continuously adapted to the developing system. The data can be exported from the system for further analyses.

Teacher-Assembled Tests

Teachers have been encouraged to use objective assessment instruments since the very beginning of educational testing; however, most tests available for classroom assessment are summative tests. Such tests are difficult to adapt to the actual needs of a class, not to mention individual students. Another option is teacher-made tests, but the time and resources needed to prepare and score them hinder practical use. The teacher-assembled tests in eDia fill this gap. Participating teachers are granted access to the item banks, so they can assemble tests out of available tasks. These tests can then be administered to individual students, a group of students or an entire class, with the results made available immediately after testing. Models for the co-existence of centrally initiated tests and teachers' assessment are under development. The current model is that central assessments serve a screening function, while teacher-initiated tests are mostly used for formative and diagnostic purposes if needed. Further options are being explored, e.g., automated recommendations for testing based on previous assessment results.

Feedback

At present, there are two basic forms of feedback. One is the immediate feedback students receive right after the test has been completed in the form of percentage of total score of a particular test. Another form is contextualized information based on normative reference data, available only after the central assessments. After the general assessments, both students and teachers receive detailed information about the results for each assessment dimension. Students may download a PDF file with a detailed description of the content of the assessment and their own achievement compared to the national norm and class mean. Teachers receive similar information on their students individually in each dimension as well as a comprehensive, contextualized picture of their class, comparing it to other members of the same age group in the entire school, school district, region, and country. This feedback is provided in graphic form as well to help teachers comprehend and use the data.

Scaling and Setting Norms

An IRT model is used to establish assessment scales. There are nine distinct scales in the eDia system as they are defined in the assessment framework; each one is developed separately. Establishing normative scales is a long process, one which requires several steps in the case of the eDia system. The results of the end-of-year assessments are used to establish the scales. In the first step, separate norms are defined for the different grades, with the mean for a grade set for 500 with a SD of 100. This phase has already been completed, and the 54 (6 grades × 3 domains × 3 dimensions) reference scales have been established.

The next step is to devise developmental scales with vertical scaling of the data, linking the achievement of the different grades. This can be done easily with a psychological dimension, where a more or less continuous development can be assumed. As cognitive development is stimulated by out-of-school experiences as well, there may be large differences within a given cohort; some students' achievement may be closer to the mean for a different cohort. Thus, linking the grades causes no difficulties. These considerations are only partially appropriate for the application dimensions, while the disciplinary dimensions are based on the material taught. Therefore, students in a particular grade may only be offered tasks from earlier grades, but not from later ones. Due to these complications, the first vertical scales for the psychological dimensions have already been prepared (see Molnár and Csapó, submitted), but vertical scaling in the other two dimensions requires more sophisticated statistical procedures (e.g., multidimensional IRT).

Finally, longitudinal scales will also be devised, making it possible to monitor student progress and to observe how they progress within a given period, compared to his/her previous and others' mean change. Developing such scales requires even more care and time and is especially difficult because collecting longitudinal data from the period covered by eDia takes at least 5 years, while the social and contextual conditions are also rapidly changing in the meantime. On the other hand, eDia does not provide high-stakes testing, nor is producing trend data a requirement. Thus, it can be flexible in establishing normative scales. Whatever the means used for scaling, scale development should also serve the formative, diagnostic function of the system.

Novel Item Formats for Improving the Quality of Testing

Quality of testing can be defined in terms of validity (including predictive and diagnostic validity), reliability, and objectivity. In this section, we show how new item formats made possible by technology can improve the quality of testing. A number of media effect studies have been carried out in past decades to explore most aspects of assessments. The quality of TBA is usually compared to paper-and-pencil or face-to-face testing, so we also compare the eDia items to these traditional testing modes. Technology offers numerous new options both in presenting stimuli and in capturing students' responses that are not possible through traditional testing modes; in addition, technology improves objectivity and validity significantly (for a detailed discussion of technological issues, see Csapó et al., 2012).

New Forms of Stimuli

Use of technology expands the possibilities of creating more life-like situations and using more authentic stimuli. There are three ways to develop computer-based tests, tasks, and items. First, tests/tasks/items can be prepared according to traditional approaches with designs based on paper-and-pencil techniques. Texts, static images, schematic figures, and graphs are also available on paper, but their richness and variety represent an added value of TBA. We call these kinds of computer-based tasks first-generation tasks (Molnár et al., 2017). Second-generation tests contain tasks with new formats, including multimedia (e.g., animation, video and audio), constructed response, automatic item generation, and automatic scoring tests (Pachler et al., 2010), thus increasing the level of authenticity and the power of assessment. These types of tasks cannot be administered in paper-and-pencil format. Finally, third-generation tests dramatically increase the level of reality and the number of ways students can demonstrate their skills as they allow students to interact with complex scenarios (e.g., complex problem-solving items in the MicroDYN approach), simulations (html documents to imitate a closed internet environment), situations (e.g., GeoGebra elements), and dynamically changing items and/or to collaborate online with other students to solve dynamically changing, interactive problem-solving items. All of these options are implemented and available for item development in the eDia system.

Any kind of multimedia, animation, video, voice, etc. provides authentic content, improves validity, and serves specific functions. Special accommodations can be embedded into technology-based tests; for example, validity of test results can be enhanced by providing instructions both in an on-screen written form and with a pre-recorded voice, thereby preventing failures caused by students' reading difficulties. Thus, in the eDia system, students in Grades 1–3 can listen to instructions on headphones while the tests are being administered. It is also possible to standardize the test environment by controlling the presentation of information in different ways (e.g., timing and a given number of repetitions).

New Forms for Response Capture

Use of technology changes not only the forms of stimuli but also those of response capture. In the traditional test environment, response capture happened basically by circling, ticking, X-ing, underlining or writing letters, numbers, words or sentences. The TBA environment expands these options, but this expansion strongly depends on the technology used. There are different possibilities for response capture in the case of a tablet or a desktop computer. The eDia system is prepared for both. However, as the keyboard and mouse are used for input in most Hungarian schools, the eDia task responses are optimized for them.

The TBA environment makes it possible to expand the possibilities of manipulation with task elements and to realize the following forms of response capture with a mouse: (1) clicking on form elements (radio button and checkbox), (2) using a drop-down menu, (3) clicking on pictures or parts of pictures, (4) clicking on texts or parts of texts, (5) coloring shapes or pictures or parts of them by clicking, (6) sequencing by ordering mouse clicks, (7) connecting two task elements with lines or arrows, (8) constructing answers with on-screen manipulations with drag-and-drop letters, words, sentences, numbers, shapes, pictures, voices, sounds, animations, simulations, etc., that is, all kinds of task elements, and (9) using sliders and functions or other changeable and interactive task elements. Other possibilities are available with the keyboard, such as typing letters, numbers, and words. Logging and analyzing log data by measuring response time, mouse movement, and navigation sequence to describe the activity of the students during testing can also contribute to more elaborated feedback; however, further studies are required to explore how to use these methods more effectively. All these possibilities for logging students' activities while they respond to items are available in the eDia system.

Complex Item Formats: Interactivity and Simulation

The eDia system was prepared to administer third-generation tests. The MicroDYN-based assessment of problem-solving (Greiff and Funke, 2009; Greiff et al., 2013; Molnár and Csapó, 2018) is available with a large number of items. One of the benefits of MicroDYN is that it allows various independent and dependent variables, and different connections may be defined between them for the simulated systems. The difficulty level of the task may thus easily be changed. A further expansion of this conception is the assessment of collaborative problem-solving. It makes it possible to use a real human-human scenario during data collection (Pásztor-Kovács et al., 2018). This allows more social interaction, compared to the PISA 2015 collaborative problem-solving assessment, which used human-agent interaction (OECD, 2017). Further simulation-based items were used on an ICT literacy test (Tongori, 2018). These complex item formats have been used for assessments beyond the diagnostic system and for experimentation and research, and these experiences will also be applied to the diagnostic assessments.

BEYOND DIAGNOSTIC ASSESSMENT: eDia AS A RESEARCH INSTRUMENT

Beyond its main purpose of providing diagnostic assessments, the eDia platform has been used in a number of other domains and in research projects as well. In this section, we review the research in which data were collected by eDia.

Further Assessment Domains Implemented in eDia

At present, there are over 20 further domains (called minor domains) for which tests or test batteries are implemented on the eDia platform. The principle in general is that different tests are prepared for the different age groups linked with anchor items.

Supporting the kindergarten-school transition with assessment instruments is one of the current extensions of the eDia. First, the DIFER test battery, a broadly used face-to-face instrument, was digitized, and then the traditional and online delivery methods were compared. Results from the media effect study indicated that the two versions (face-to-face vs. online) were equivalent and that the digitized version was not only more convenient to use, but the objectivity and reliability had also improved on some subtests (Csapó et al., 2014). Based on these experiences, a new school readiness test battery has been developed and optimized for online assessment, which can be used in kindergarten with tablets (Csapó et al., 2017, 2018).

Several instruments were devised for assessments of curricular areas beyond the three major domains. The media effect on composing skills was studied with primary school students (Nagy, 2015). A test of musical abilities used pre-recorded sound stimuli for melody and rhythm (Asztalos and Csapó, 2017). Several tests were prepared for English and German as a Second Language (reading, listening, and vocabulary), while the TBA made it possible to use authentic voice recordings to assess listening skills (Vígh et al., 2015; Nikolov and Csapó, 2017, 2018; Habók and Magyar, 2018a, 2019). Assessments of visual skills benefitted especially from the possibilities of rich illustrations (Kárpáti et al., 2015). Online tests have also been prepared for cross-curricular competencies, such as learning to learn (Habók, 2015; Vainikainen et al., 2015), health literacy (Nagy et al., 2015), financial literacy (Tóth, 2015), ICT literacy (Molnár et al., 2015b), and civic competencies (Kinyó, 2015).

Assessment of a variety of reasoning skills is embedded in the mathematics and science psychology dimension, mostly operational reasoning skills. However, there are some skills that play a distinct role in learning and cognitive development; therefore, comprehensive instruments have been prepared to assess them. Inductive reasoning is one of the most frequently assessed higher-order thinking skills, and several inductive reasoning tests have been developed for the eDia as well. First, a widely used paper-and-pencil inductive reasoning test (verbal and numerical analogies, and number series, see Csapó, 1997) was migrated to the digital platform (Csapó et al., 2009). Later, other tests based on Klauer's model (see, e.g., Klauer and Phye, 2008) were prepared (Molnár et al., 2013) and used in a number of national and international projects. Specific item formats were developed to assess dynamic problem-solving (the MicryDYN base, see Molnár and Pásztor-Kovács, 2015; Csapó and Molnár, 2017a), collaborative problem-solving (e.g., interactivity and communicating with pre-defined messages, see Pásztor-Kovács et al., 2018), creativity (divergent thinking and a program for counting rare solutions, see Pásztor et al., 2015), and combinatorial reasoning (drag-and-drop to combine elements and an algorithm to distinguish valid and invalid combinations, see Pásztor et al., 2015).

Tests, test batteries, and questionnaires beyond the cognitive domain are also implemented through eDia. Some of them are essential for successful learning, but because of the lack of easy-to-use instruments, they are rarely assessed. Motivation is one such affective attribute, and a related mastery motivation questionnaire is available on eDia (Józsa et al., 2015; Zsolnai and Kasik, 2015), as well as a self-regulated foreign language learning strategy questionnaire (Habók and Magyar, 2018b). The PISA 2020 learning strategy questionnaire (Artelt et al., 2003) has also been implemented and used in several projects (e.g., Csapó and Molnár, 2017a). Experimenting with the assessment of further affective and social skills is also in progress (e.g., Zsolnai and Kasik, 2015).

The eDia platform has been used in higher education. For example, in 2015, the University of Szeged introduced an assessment system to explore how well incoming students are prepared for university studies. In the first year, six tests were administered through eDia: Hungarian language and literature (with a strong reading comprehension component), mathematics, history, science and English as a foreign language as well as a dynamic problem-solving test (Csapó and Molnár, 2017a). Since then, the system has evolved further (Molnár and Csapó, 2019b).

Applications of eDia in International Assessments; Comparative Studies

The eDia system has been used for research within international collaborative projects carried out by the University of Szeged Centre for Research on Learning and Instruction and supports investigations by PhD students at the Doctoral School of Education at the same university. In this section, we review some results of these efforts, highlighting new opportunities for educational research offered by the online assessment.

In Finland, the Centre for Educational Assessment, University of Helsinki, cooperates with Vantaa city schools in using tablets in everyday teaching and learning processes. Within the framework of this project, Hungarian tests were translated into Finnish and assessments were carried out in both countries using the same instruments, with the tests delivered from the University of Szeged servers (Hotulainen et al., 2018; Pásztor et al., 2018). The first results may indicate the impact of frequent testing, but further studies would be required to uncover the mechanisms.

The tests for assessing thinking skills implemented in the eDia have been used in several international studies. The knowledge acquisition phase of dynamic problem-solving involves two further skills, combinatorial reasoning (systematically combining possible values of independent variables) and inductive reasoning (rule induction and generalizing the experience of interactions). The relationships of these skills were explored; the dynamic problem-solving tests, together with combinatorial and inductive reasoning tests were translated into Chinese and were administered to Chinese students. The results indicated a stronger impact of combinatorial reasoning than that of inductive reasoning (Wu and Molnár, 2018a). The relationship between problem-solving, creativity, inductive reasoning, and working memory was explored in a similar study (Wu and Molnár, 2018b). In Namibia, the relationship between scientific reasoning and motivation to learn science was examined (Kambeyo et al., 2017) as well as the possibilities of online assessment of scientific inquiry skills. These studies indicated that online assessment is feasible even with a modest school infrastructure.

Another set of studies was completed on learning foreign languages in three countries, Mongolia (Ragchaa, 2017), Kazakhstan (Akhmetova and Csapó, 2018), and Azerbaijan (Karimova and Csapó, 2018), where the two most frequently studied foreign languages are English and Russian. Thus, these countries offer different contexts and sets of conditions than those of Hungary, where the main foreign languages are English and German (see, e.g., Nikolov and Csapó, 2018). Another difference is that these countries use the Cyrillic alphabet. Several research questions were explored in these studies on learning foreign languages with eDia-based instruments, including the development of receptive skills, self-concept and learning strategies.

Assessment Platform for the Hungarian Educational Longitudinal Program

The Hungarian Educational Longitudinal Program (HELP) was launched in 2003 and is maintained by the SZTE-MTA Research Group on the Development of Competencies (Csapó, 2007). A new cohort (a nationally representative sample of approx. 6,000 students) is added to the program every 4 years, with students being monitored from the beginning of schooling to the end of compulsory education. Data collection has focused on three main domains, reading, mathematics, and science, and data are systematically collected on a number of cognitive, affective, and contextual variables. The online assessment has been gradually introduced to the data collection effort (e.g., languages have been tested online, see Nikolov and Csapó, 2018), with the cohort that entered school in 2015 having been exclusively assessed with the eDia instruments. The benefit of longitudinal research from the perspective of developing the diagnostic system is that it offers a nationally representative sample for scale development and for determining the predictive power of certain instruments (e.g., school readiness tests, see Csapó et al., 2018).

DISCUSSION AND CONCLUSIONS

Practical Relevance and Limitations of the Online Assessment

Systematic feedback is a basic condition for the operation and development of any complex system and providing students and teachers with an inexpensive, easy-to-use, valid, and reliable assessment system may significantly contribute to solving certain crucial problems of education today. Making it possible to measure the different dimensions of learning separately, especially the mostly hidden psychological dimension, i.e., thinking and cognitive development may support meaningful learning and a deeper conceptual understanding. (Empirical studies concerning these assumptions are in progress; see also Molnár and Csapó, 2019a).

Teachers see the differences between their students and realize if some of their students fail, but without proper instruments teachers cannot determine the nature and magnitude of the differences with precision. Diagnostic assessments support the personalization of learning, adjusting teaching to students' personal needs. Teachers routinely use certain types of formative assessment (mostly based on their subjective observation), and we may assume that with better instruments they will teach better. However, we may not assume that they will be able to fully exploit the potential of online diagnostic assessments; they need training to empower them. Several training programs (from one-day introductory workshops to two-year training of assessment experts) are available within the framework of the project. Ideally, the teacher-training component is an in-service adaptation of research-based teacher education (see, e.g., Munthe and Rogne, 2015).

As there is a growing concern among teachers about high-stakes testing and the use of its results for accountability (Tóth, 2011), monitoring their views on diagnostic assessment will be an important task. An indicator of acceptance of eDia is that teachers and schools have been participating in the assessments voluntarily, with informal communication confirming its acceptance as well. Formal surveys will be needed to gain a better understanding of teachers' opinions.

Finally, we have to emphasize that an assessment instrument alone does not improve the quality of learning; its practical impact depends on how the information it provides is used to change teaching and learning processes. To better use the power of feedback, the conception of classroom teaching should basically be changed; there is a need for new models of teaching and learning, where students' individual needs are better served. Such models have existed for decades, but the lack of appropriate tools has hindered large-scale use. In the most general terms, Mastery Learning is one such model, which, supported with online pre-tests and post-tests, may gain a new impetus (Csapó and Molnár, 2017b). There are also several promising new models which stress the role of regular feedback and use of assessment data made possible by TBA, e.g., data-based teaching (Datnow and Hubbard, 2016) and assessment-powered teaching (Sindelar, 2010). Experience in the areas of computer aided-instruction and tutoring systems (Kulik and Fletcher, 2016; Chauhan, 2017) may be used, especially in stimulating students' development in the psychological dimensions when diagnostic assessments indicate the need for such intervention.

Further Research Prospects

Regular diagnostic assessments generate large databases and render it possible to make further sophisticated use of those that have already been started in other areas (see research on the "data revolution" and "big data"). Educational data mining and process mining have already produced results applicable in practice as well (Tóth et al., 2017). Certain methods developed within the paradigm of learning analytics may also be used to process databases produced by diagnostic assessments as well.

Log file analysis is the easiest and most appropriate new method for using new types of assessment data (metadata and log data). An easily recordable and already routinely used piece of information is the time students spend on certain activities when completing online tasks; time-on-task analyses, among other methods, may indicate students' attention and motivation. Some item types (combinatorial reasoning task enumerations, MicroDYN items and collaborative problem-solving activities) allow the recording of more detailed information on students' reasoning. Some analyses (e.g., latent class analyses) using data collected with eDia have already been conducted (Greiff et al., 2018; Molnár and Csapó, 2018), but further research is needed to find ways to make practical use of these results, adding new analytical modules to the eDia platform, creating new, log data-based indicators and supporting students' cognitive development in the long run.

AUTHOR CONTRIBUTIONS

Both of the authors, BC and GM, certify that they have participated sufficiently in the study to take responsibility for the content, including writing and final approval of the manuscript. Each author agrees to be accountable for all aspects of the paper.

FUNDING

Preparation of this article was funded by the Hungarian Academy of Sciences through the MTA-SZTE Research Group on the Development of Competencies and by OTKA K115497.

ACKNOWLEDGMENTS

The authors wish to thank all the developers of eDia for their contributions, including framework developers, programmers, item writers and support staff at the Centre for Research on Learning and Instruction. Special thanks to the teachers at our partner schools who have been encouraging us and using the diagnostic system in their everyday work.

REFERENCES

Ainsworth, L., and Viegut, D. (2006). *Common formative assessments. How to connect standards-based instruction and assessment*. (Thousand Oaks, CA: Corwin Press).

Akhmetova, A., and Csapó, B. (2018). Development of reading skills of 6th and 8th graders in English, Kazakh, and Russian from the perspective of young learners' backgrounds in Pavlodar. *XVIII. Országos Neveléstudományi Konferencia*; Budapest.

Ansari, D., and Coch, D. (2006). Bridges over troubled water: education and cognitive neuroscience. *Trends Cogn. Sci.* 10, 146–151. doi: 10.1016/j.tics.2006.02.007

Artelt, C., Baumert, J., Julius-McElvani, N., and Peschar, J. (2003). *Learners for life. Student approaches to learning. Results from PISA 2000*. (Paris: OECD).

Asztalos, K., and Csapó, B. (2017). Development of musical abilities: cross-sectional computer-based assessments in educational contexts. *Psychol. Music* 45, 682–698. doi: 10.1177/0305735616678055

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. (New York: Holt, Rinehart and Winston).

Avella, J. T., Kebritchi, M., Nunn, S. G., and Kanai, T. (2016). Learning analytics methods, benefits, and challenges in higher education: a systematic literature review. *Online Learn.* 20, 13–29.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): a preliminary theory of action for summative and formative assessment. *Measurement* 8, 70–91. doi: 10.1080/15366367.2010.508686

Bennett, R. E. (2011). Formative assessment: a critical review. *Assess. Educ. Princ. Policy Pract.* 18, 5–25. doi: 10.1080/0969594x.2010.513678

Bennett, R. E., and Gitomer, D. H. (2009). "Transforming K-12 assessment" in *Assessment issues of the 21st century*. eds. C. Wyatt-Smith and J. Cumming (New York, NY: Springer), 43–61.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., et al. (2012). "Defining twenty-first century skills" in *Assessment and teaching of 21st century skills*. eds. P. Griffin, B. McGaw, and E. Care (New York: Springer), 17–66.

Black, P., Harrison, C., Lee, C., Marshall, B., and William, D. (2003). *Assessment for learning. Putting it into practice*. (Berkshire: Open University Press).

Black, P., Harrison, C., Lee, C., Marshall, B., and Wiliam, D. (2004). Working inside the black box: assessment for learning in the classroom. *Phi Delta Kappan* 86, 8–21. doi: 10.1177/003172170408600105

Black, P., and Wiliam, D. (1998a). *Inside the black box: Raising standards through classroom assessment*. (London, UK: King's College).

Black, P., and Wiliam, D. (1998b). Assessment and classroom learning. *Assess. Educ. Princ. Policy Pract.* 5, 7–74. doi: 10.1080/0969595980050102

Brendefur, J. L., Johnson, E. S., Thiede, K. W., Strother, S., and Severson, H. H. (2018). Developing a multi-dimensional early elementary mathematics screener and diagnostic tool: the primary mathematics assessment. *Early Childhood Educ. J.* 46, 153–157. doi: 10.1007/s10643-017-0854-x

Brown, J., Hinze, S., and Pellegrino, J. W. (2008). "Technology and formative assessment" in *21st century education. Vol 2. Technology*. ed. T. Good (Thousand Oaks, CA: Sage), 245–255.

Chauhan, S. (2017). A meta-analysis of the impact of technology on learning effectiveness of elementary students. *Comput. Educ.* 105, 14–30. doi: 10.1016/j.compedu.2016.11.005

Chi, M., VanLehn, K., Litman, D., and Jordan, P. (2010). "Inducing effective pedagogical strategies using learning context features" in *User modeling, adaptation and personalization: 18th international conference, UMAP 2010*. eds. P. De Bra, A. Kobsa, and D. Chin (Heidelberg: Springer), 147–158.

Clarke, S. (2001). *Unlocking formative assessment. Practical strategies for enhancing pupils learning in primary classroom*. (London: Hodder Arnold).

Clarke, S. (2005). *Formative assessment in action. Weaving the elements together*. (London: Hodder Murray).

Conejo, R., Guzmán, E., Millán, E., Trella, M., Pérez-De-La-Cruz, J. L., and Ríos, A. (2004). SIETTE: a web-based tool for adaptive testing. *Int. J. Artif. Intell. Educ.* 14, 29–61.

Conejo, R., Millán, E., Pérez-de-la-Cruz, J. L., and Trella, M. (2000). "An empirical approach to on-line learning in SIETTE" in *Intelligent tutoring systems. Proceedings of the 5th international conference on intelligent tutoring systems*. eds. G. Gauthier, C. Frasson, and K. VanLehn (Berlin, Heidelberg: Springer), 605–614.

Csapó, B. (1997). The development of inductive reasoning: cross-sectional measurements in an educational context. *Int. J. Behav. Dev.* 20, 609–626.

Csapó, B. (2004). "Knowledge and competencies" in *The integrated person. How curriculum development relates to new competencies*. ed. J. Letschert (CIDREE: Enschede), 35–49.

Csapó, B. (2007). Hosszmetszeti felmérések iskolai kontextusban – az első átfogó magyar iskolai longitudinális kutatási program elméleti és módszertani keretei [Longitudinal assessments in school context – theoretical and methodological frames of the first large-scale school-related longitudinal program in Hungary]. *Magyar Pedagógia* 107, 321–355.

Csapó, B. (2010). Goals of learning and the organization of knowledge. *Z. Pädagog.* 2010(Suppl. 56), 12–27.

Csapó, B., Ainley, J., Bennett, R., Latour, T., and Law, N. (2012). "Technological issues of computer-based assessment of 21st century skills" in *Assessment and teaching of 21st century skills*. eds. P. Griffin, B. McGaw, and E. Care (New York: Springer), 143–230.

Csapó, B., and Csépe, V. (eds.) (2012). *Framework for diagnostic assessment of reading*. (Budapest: Nemzeti Tankönyvkiadó).

Csapó, B., Csíkos, C., and Molnár, G. (eds.) (2015a). *A matematikai tudás diagnosztikus értékelésének tartalmi keretei [Framework for online diagnostic assessment of mathematics knowledge]*. (Budapest: Oktatáskutató és Fejlesztő Intézet).

Csapó, B., and Funke, J. (eds.) (2017). *The nature of problem solving. Using research to inspire 21st century learning*. (Paris: OECD).

Csapó, B., Hódi, Á., Kiss, R., Pásztor, A., Rausch, A., and Molnár, G. (2017). Developing online diagnostic instruments for assessing pupils' skills at the beginning of schooling. *Paper presented at the 17th biennial conference for research on learning and instruction*; Tampere.

Csapó, B., Korom, E., and Molnár, G. (eds.) (2015b). *A természettudományi tudás diagnosztikus értékelésének tartalmi keretei [Framework for online diagnostic assessment of science knowledge]*. (Budapest: Oktatáskutató és Fejlesztő Intézet).

Csapó, B., and Molnár, G. (2017a). Potential for assessing dynamic problem-solving at the beginning of higher education studies. *Front. Psychol.* 8, 1–12. doi: 10.3389/fpsyg.2017.02022

Csapó, B., and Molnár, G. (2017b). "Assessment-based, personalized learning in primary education" in *Knowledge management in the 21st century: Resilience, creativity and co-creation. Proceedings of IFKAD 2017*. eds. J. C. Spender, G. Schiuma, and T. Gavrilova (St. Petersburg: St. Petersburg University Graduate School of Management), 443–449.

Csapó, B., Molnár, G., and Nagy, J. (2014). Computer-based assessment of school-readiness and reasoning skills. *J. Educ. Psychol.* 106, 639–650. doi: 10.1037/a0035756

Csapó, B., Molnár, G., and Pásztor, A. (2018). Predictive validity of technology-based school-readiness assessments. *Paper presented at the 9th biennial conference of EARLI SIG 1, assessment and evaluation: Assessment & learning analytics*; Helsinki.

Csapó, B., Molnár, G., Tóth, R., and K., (2009). "Comparing paper-and-pencil and online assessment of reasoning skills. A pilot study for introducing electronic testing in large-scale assessment in Hungary" in *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing*. eds. F. Scheuermann and J. Björnsson (Luxemburg: Office for Official Publications of the European Communities), 120–125.

Csapó, B., and Pásztor, A. (2015). "A kombinatív képesség fejlődésének mérése online tesztekkel [Assessment of the development of combinative ability with online tests]" in *Online diagnosztikus mérések az iskola kezdő szakaszában [Online diagnostic assessments in the beginning phase of schooling]*. eds. B. Csapó and A. Zsolnai (Budapest: Oktatáskutató és Fejlesztő Intézet), 367–386.

Csapó, B., Steklács, J., and Molnár, G. (eds.) (2015c). *Az olvasás-szövegértés online diagnosztikus értékelésének tartalmi keretei [Framework for online diagnostic assessment of reading comprehension]*. (Budapest: Oktatáskutató és Fejlesztő Intézet).

Csapó, B., and Szabó, G. (eds.) (2012). *Framework for diagnostic assessment of science*. (Budapest: Nemzeti Tankönyvkiadó).

Csapó, B., and Szendrei, M. (eds.) (2011). *Framework for diagnostic assessment of mathematics*. (Budapest: Nemzeti Tankönyvkiadó).

Csíkszentmihályi, M. (2000). *Beyond boredom and anxiety*. (San Francisco: Jossey-Bass). (Original work published 1975).

Darling-Hammond, D. (2012). "Policy frameworks for new assessments" in *Assessment and teaching of 21st century skills*. eds. P. Griffin, B. McGaw, and E. Care (New York: Springer), 301–339.

Datnow, A., and Hubbard, L. (2016). Teacher capacity for and beliefs about data-driven decision making: a literature review of international research. *J. Educ. Chang.* 17, 7–28. doi: 10.1007/s10833-015-9264-2

Farcot, M., and Latour, T. (2009). "Transitioning to computer-based assessments: a question of costs" in *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing*. eds. F. Scheuermann and J. Björnsson (Luxemburg: Office for Official Publications of the European Communities), 108–116.

Feng, M., and Heffernan, N. T. (2005). Informing teachers live about student learning: reporting in the assistment system. *Technol. Instr. Cogn. Learn.* 3, 1–14.

Feng, M., Heffernan, N. T., and Koedinger, K. R. (2009). Addressing the assessment challenge in an online system that tutors as it assesses. *User Model. User-Adapt. Interact.* 19, 243–266. doi: 10.1007/s11257-009-9063-7

Filderman, M. J., Toste, J. R., Didion, L. A., Peng, P., and Clemens, N. H. (2018). Data-based decision making in reading interventions: a synthesis and meta-analysis of the effects for struggling readers. *J. Spec. Educ.* 52, 174–187. doi: 10.1177/0022466918790001

Good, R. (2011). Formative use of assessment information: it's a process, so let's say what we mean. *Pract. Assess. Res. Eval.* 16, 1–6.

Greiff, S., and Funke, J. (2009). "Measuring complex problem solving: the MicroDYN approach" in *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing*. eds. F. Scheuermann and J. Björnsson (Luxemburg: Office for Official Publications of the European Communities), 157–163.

Greiff, S., Molnár, G., Martin, R., Zimmermann, J., and Csapó, B. (2018). Students' exploration strategies in computer-simulated complex problem environments: a latent class approach. *Comput. Educ.* 126, 248–263. doi: 10.1016/j.compedu.2018.07.013

Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F., and Funke, J. (2013). Computer-based assessment of complex problem solving: concept, implementation, and application. *Educ. Technol. Res. Dev.* 61, 407–421. doi: 10.1007/s11423-013-9301-x

Griffin, P., and Care, E. (eds.) (2015). *Assessment and teaching of 21st century skills: Methods and approach*. (New York: Springer).

Habók, A. (2015). "A tanulás tanulásának vizsgálata online környezetben [Exploration of learning to learn in an online environment]" in *Online diagnosztikus mérések az iskola kezdő szakaszában [Online diagnostic assessments in the beginning phase of schooling]*. eds. B. Csapó and A. Zsolnai (Budapest: Oktatáskutató és Fejlesztő Intézet), 179–198.

Habók, A., and Magyar, A. (2018a). The effect of language learning strategies on proficiency, attitudes and school achievement. *Front. Psychol.* 8:2358. doi: 10.3389/fpsyg.2017.02358

Habók, A., and Magyar, A. (2018b). Validation of a self-regulated foreign language learning strategy questionnaire through multidimensional modelling. *Front. Psychol.* 9:1388. doi: 10.3389/fpsyg.2018.01388

Habók, A., and Magyar, A. (2019). The effects of EFL reading comprehension and certain learning related factors on EFL learners' reading strategy use. *Cogent Educ.* 6, 1–19. doi: 10.1080/2331186x.2019.1616522

Haldane, S. (2009). "Delivery platforms for national and international computer-based surveys: history, issues and current status" in *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing*. eds. F. Scheuermann and J. Björnsson (Luxemburg: Office for Official Publications of the European Communities), 63–67.

Halldórsson, A. M., McKelvie, P., and Björnsson, J. K. (2009). "Are Icelandic boys really better on computerized tests than conventional ones? Interaction between gender, test modality and test performance" in *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing*. eds. F. Scheuermann and J. Björnsson (Luxemburg: Office for Official Publications of the European Communities), 178–193.

Hattie, J. A., and Brown, G. T. (2007). Technology for school-based assessment and assessment for learning: development principles from New Zealand. *J. Educ. Technol. Syst.* 36, 189–201. doi: 10.2190/et.36.2.g

Heitink, M. C., Van der Kleij, F. M., Veldkamp, B. P., Schildkamp, K., and Kippers, W. B. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educ. Res. Rev.* 17, 50–62. doi: 10.1016/j.edurev.2015.12.002

Hotulainen, R., Pásztor, A., Kupiainen, S., Molnár, G., and Csapó, B. (2018). Entering school with equal skills? A two-country comparison of early inductive reasoning. *Paper presented at the 9th biennial conference of EARLI SIG 1, assessment and evaluation: Assessment & learning analytics*; Helsinki.

Józsa, K., Hricsovinyi, J., and Szenczi, B. (2015). "Számítógép-alapú elsajátítási motiváció kérdőívek validitása és reliabilitása [Validity and reliability of computer-based mastery motivation questionnaires]" in *Online diagnosztikus mérések az iskola kezdő szakaszában [Online diagnostic assessments in the beginning phase of schooling]*. eds. B. Csapó and A. Zsolnai (Budapest: Oktatáskutató és Fejlesztő Intézet), 123–146.

Kambeyo, L. (2017). The possibilities of assessing students' scientific inquiry skills abilities using an online instrument: a small-scale study in the Omusati Region, Namibia. *Eur. J. Educ. Sci.* 4, 1–21. doi: 10.19044/ejes.v4no2a1

Kambeyo, L., Pásztor, A., Korom, E. B., Németh, M., and Csapó, B. (2017). Online assessment of scientific reasoning and motivation to learn science: a pilot study in Namibia. *Paper presented at the 17th biennial conference for research on learning and instruction*; Tampere.

Karimova, K., and Csapó, B. (2018). Listening and reading self-concepts in the English and Russian languages. *XVIII. Országos Neveléstudományi Konferencia*; Budapest.

Kárpáti, A., Babály, B., and Simon, T. (2015). "A vizuális képességrendszer elemeinek értékelése: térszemlélet és kepi kommunikáció. [Evaluation of the elements of the system of visual skills: spatial representation and pictorial communication]" in *Online diagnosztikus mérések az iskola kezdő szakaszában [Online diagnostic assessments in the beginning phase of schooling]*. eds. B. Csapó, and A. Zsolnai (Budapest: Oktatáskutató és Fejlesztő Intézet), 35–69.

Kinyó, L. (2015). "A társadalmi és állampolgári ismeretek online vizsgálata [Online assessment of civic competencies and knowledge about society]" in *Online diagnosztikus mérések az iskola kezdő szakaszában [Online diagnostic assessments in the beginning phase of schooling]*. eds. B. Csapó and A. Zsolnai (Budapest: Oktatáskutató és Fejlesztő Intézet), 97–121.

Klauer, K. J., and Phye, G. D. (2008). Inductive reasoning: a training approach. *Rev. Educ. Res.* 78, 85–123. doi: 10.3102/0034654307313402

Koretz, D. (2018). Moving beyond the failure of test-based accountability. *Am. Educ.* 41, 22–26.

Kozma, R. (2009). "Transforming education: assessing and teaching 21st century skills. Assessment call to action" in *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing*. eds. F. Scheuermann and J. Björnsson (Luxemburg: Office for Official Publications of the European Communities), 13–23.

Kulik, J. A., and Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: a meta-analytic review. *Rev. Educ. Res.* 86, 42–78. doi: 10.3102/0034654315581420

Lee, M.-K. (2009). "CBAS in Korea: experiences, results and challenges" in *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing*. eds. F. Scheuermann, and J. Björnsson (Luxemburg: Office for Official Publications of the European Communities), 194–200.

Leighton, J. P., and Gierl, M. J. (eds.) (2007). *Cognitive diagnostic assessment for education: Theory and applications*. (New York, NY, US: Cambridge University Press).

McLachlan, C., Fleer, M., and Edwards, S. (2018). *Early childhood curriculum: Planning, assessment and implementation*: (New York, NY, US: Cambridge University Press).

Molnár, G., and Csapó, B. (2018). The efficacy and development of students' problem-solving strategies during compulsory schooling: logfile analyses. *Front. Psychol.* 9:302. doi: 10.3389/fpsyg.2018.00302

Molnár, G., and Csapó, B. (2019a). Making the psychological dimension of learning visible: using technology-based assessment to monitor students' cognitive development. *Front. Psychol.* 10:1368. doi: 10.3389/fpsyg.2019.01368

Molnár, G., and Csapó, B. (2019b). A felsőoktatási tanulmányi alkalmasság értékelésére kidolgozott rendszer a Szegedi Tudományegyetemen: elméleti keretek és mérési eredmények [The system developed for the assessment of preparedness for higher educational studies at the University of Szeged: theoretical frameworks and measurement results]. *Education* (in press).

Molnár, G., Greiff, S., and Csapó, B. (2013). Inductive reasoning, domain specific and complex problem solving: relations and development. *Think. Skills Creat.* 9, 35–45. doi: 10.1016/j.tsc.2013.03.002

Molnár, G., Greiff, S., Wüstenberg, S., and Fischer, A. (2017). "Empirical study of computer based assessment of domain-general dynamic problem solving skills" in *The nature of problem solving*. eds. B. Csapó, and J. Funke (Paris: OECD), 123–143.

Molnár, G., Makay, G., and Ancsin, G. (2018). *Feladat- és tesztszerkesztés az eDia-rendszerben [Task and test development in the eDia system]*. (Szeged: Szegedi Tudományegyetem Oktatáselméleti Kutatócsoport).

Molnár, G., Papp, Z., Makay, G., and Ancsin, G. (2015a). *eDia 2.3 Online mérési platform – feladatfelviteli kézikönyv [eDia 2.3 Online assessment platform – task editing manual]*. (Szeged: Szegedi Tudományegyetem Oktatáselméleti Kutatócsoport).

Molnár, G., and Pásztor-Kovács, A. (2015a). "A problémamegoldó gondolkodás mérése online tesztkörnyezetben [Measurement of problem solving in an online assessment environment]" in *Online diagnosztikus mérések az iskola kezdő szakaszában [Online diagnostic assessments in the beginning phase of schooling]*. eds. B. Csapó and A. Zsolnai (Budapest: Oktatáskutató és Fejlesztő Intézet), 341–366.

Molnár, G., Tongori, Á., and Pluhár, Z. (2015b). "Az informatikai műveltség online mérése [Online assessment of info-communication literacy]" in *Online diagnosztikus mérések az iskola kezdő szakaszában [Online diagnostic assessments in the beginning phase of schooling]*. eds. B. Csapó and A. Zsolnai (Budapest: Oktatáskutató és Fejlesztő Intézet), 295–317.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O' Sullivan, C. Y., Arora, A., and Eberber, E. (eds.) (2005). *TIMSS 2007 assessment frameworks*. (Boston: TIMSS & PIRLS International Study Center, Boston College).

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., and Gonzalez, E. J. et al. (eds.) (2001). *Assessment frameworks and specifications 2003.* 2nd Edn. (Boston: International Study Center, Boston College).

Munthe, E., and Rogne, M. (2015). Research based teacher education. *Teach. Teach. Educ.* 46, 17–24. doi: 10.1016/j.tate.2014.10.006

Nagy, Z. (2015). "A médiahatás vizsgálata általános iskolás tanulók papíralapú és online fogalmazásain [Examination of the paper-based vs. online media effect on composing skills of primary school students]" in *Online diagnosztikus mérések az iskola kezdő szakaszában [Online diagnostic assessments in the beginning phase of schooling]*. eds. B. Csapó and A. Zsolnai (Budapest: Oktatáskutató és Fejlesztő Intézet), 225–244.

Nagy, L., Korom, E., Hódi, Á., and Németh, B. M. (2015). "Az egészségműveltség online mérése [Online assessment of health literacy]" in *Online diagnosztikus mérések az iskola kezdő szakaszában [Online diagnostic assessments in the beginning phase of schooling]*. eds. B. Csapó and A. Zsolnai (Budapest: Oktatáskutató és Fejlesztő Intézet), 147–177.

Nikolov, M., and Csapó, B. (2017). Reading abilities in three languages and their relationship to students' inductive reasoning skills and their parents' level of education. *Paper presented at the conference of the American association for applied linguistics*; Atlanta, GA.

Nikolov, M., and Csapó, B. (2018). The relationships between 8th graders' L1 and L2 readings skills, inductive reasoning and socio-economic status in early English and German as a foreign language programs. *System* 73, 48–57. doi: 10.1016/j.system.2017.11.001

OECD (1999). *Measuring student knowledge and skills. A new framework for assessment*. (Paris: OECD).

OECD (2003). *The PISA 2003 assessment framework. Mathematics, reading, science and problem solving.* (Paris: OECD).

OECD (2010). *PISA computer-based assessment of student skills in science*. (Paris: OECD).

OECD (2011). *PISA 2009 results: Students on line: Digital technologies and performance (volume VI)*. (Paris: OECD).

OECD (2013). *PISA 2012 results: What students know and can do – Student performance in mathematics, reading and science (volume I)*. (Paris: OECD Publishing).

OECD (2014). *PISA 2012 results: Creative problem solving. Students' skills in tackling real-life problems (volume V)*. (Paris: OECD).

OECD (2016). *PISA 2015 results (volume I): Excellence and equity in education.* (Paris: OECD).

OECD (2017). *PISA 2015 results (volume V): Collaborative problem solving.* (Paris: OECD).

Opfer, J. E., Nehm, R. H., and Ha, M. (2012). Cognitive foundations for science assessment design: knowing what students know about evolution. *J. Res. Sci. Teach.* 49, 744–777. doi: 10.1002/tea.21028

Pachler, N., Daly, C., Mor, Y., and Mellar, H. (2010). Formative e-assessment: practitioner cases. *Comput. Educ.* 54, 715–721. doi: 10.1016/j.compedu.2009.09.032

Pásztor, A., Kupiainen, S., Hotulainen, R., Molnár, G., and Csapó, B. (2018). Comparing Finnish and Hungarian fourth grade students' inductive reasoning skills. *Paper presented at the 9th biennial conference of EARLI SIG 1, assessment and evaluation: Assessment & learning Analytics*; Helsinki.

Pásztor, A., Molnár, G., and Csapó, B. (2015). Technology-based assessment of creativity in educational context: the case of divergent thinking and its relation to mathematical achievement. *Think. Skills Creat.* 18, 32–42. doi: 10.1016/j.tsc.2015.05.004

Pásztor-Kovács, A., Pásztor, A., and Molnár, G. (2018). Kollaboratív problémamegoldó képességet vizsgáló dinamikus teszt fejlesztése [Development of an online interactive instrument for assessing collaborative problem solving]. *Magyar Pedagógia* 118, 73–102. doi: 10.17670/MPed.2018.1.73

Pellegrino, J. W., Chudowsky, N., and R. Glaser (eds.) (2001). *Knowing what students know: The science and design of educational assessment*. (Washington, DC: National Academies Press).

Pellegrino, J. W., and Quellmalz, E. S., (2010). Perspectives on the integration of technology and assessment. *J. Res. Technol. Educ.* 43, 119–134. doi: 10.1080/15391523.2010.10782565

Perie, M., Marion, S., and Gong, B. (2009). Moving towards a comprehensive assessment system: a framework for considering interim assessments. *Educ. Meas. Issues Pract.* 28, 5–13. doi: 10.1111/j.1745-3992.2009.00149.x

Plichart, P., Jadoul, R., Vandenabeele, L., and Latour, T. (2004). TAO, A collective distributed computer-based assessment framework built on semantic web standards. *Proceedings of the international conference on Advances in intelligent Systems – Theory and application AISTA2004, in cooperation with IEEE computer society*; 2004 Nov 15–18; Luxembourg. Luxembourg.

Ragchaa, J. (2017). Mongolian students' learning strategies in mastering English receptive skills. *Paper presented at the 5th RSEP social sciences conference*; Barcelona.

Ryan, R. M., and Deci, E. L. (2000a). Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp. Educ. Psychol.* 25, 54–67. doi: 10.1006/ceps.1999.1020

Ryan, R. M., and Deci, E. L. (2000b). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* 55, 68–78. doi: 10.1037/0003-066X.55.1.68

Saeki, E., Segool, N., Pendergast, L., and von der Embse, N. (2018). The influence of test-based accountability policies on early elementary teachers: school climate, environmental stress, and teacher stress. *Psychol. Sch.* 55, 391–403. doi: 10.1002/pits.22112

Scandura, J. M. (2017). Contrasting fundamental assumptions in adaptive learning and modeling human tutors with TutorIT. *Technol. Instr. Cogn. Learn.* 10, 259–265.

Scardamalia, M., Bransford, J., Kozma, B., and Quellmalz, E. (2012). "New assessments and environments for knowledge building" in *Assessment and teaching of 21st century skills*. eds. P. Griffin, B. McGaw, and E. Care (New York: Springer), 231–300.

Scheuermann, F., and J. Björnsson (eds.) (2009). *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing*. (Luxemburg: Office for Official Publications of the European Communities).

Scheuermann, F., and Guimarães Pereira, A. (eds.) (2008). *Towards a research agenda on computer-based assessment: Challenges and needs for European educational measurement*. (Ispra: European Commission Joint Research Centre).

Sheard, M. K., and Chambers, B. (2014). A case of technology-enhanced formative assessment and achievement in primary grammar: how is quality assurance of formative assessment assured? *Stud. Educ. Eval.* 43, 14–23. doi: 10.1016/j.stueduc.2014.02.001

Sindelar, N. W. (2010). *Assessment-powered teaching*. (Thousand Oaks, CA: Corwin Press).

Slavin, R. E. (2002). Evidence-based education policies: transforming educational practice and research. *Educ. Res.* 31, 15–21. doi: 10.3102/0013189X031007015

Sørensen, H., and Andersen, A. M. (2009). "How did Danish students solve the PSA CBAS items?" in *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing*. eds. F. Scheuermann and J. Björnsson (Luxemburg: Office for Official Publications of the European Communities), 201–208.

Tongori, Á. (2018). Measuring ICT literacy among grade 5–11 students: Confidence in accessing information. Unpublished doctoral dissertation. Szeged: University of Szeged.

Tóth, E. (2011). Pedagógusok nézetei a tanulói teljesítménymérésekről [Teachers' views of learner assessment programmes]. *Magyar Pedagógia* 111, 225–249.

Tóth, E. (2015). "A gazdasági műveltség diagnosztikus mérésének lehetőségei online környezetben [The possibilities of assessing financial literacy in an online environment]" in *Online diagnosztikus mérések az iskola kezdő szakaszában [Online diagnostic assessments in the beginning phase of schooling]*. eds. B. Csapó, and A. Zsolnai (Budapest: Oktatáskutató és Fejlesztő Intézet), 269–293.

Tóth, K., Rölke, H., Goldhammer, F., and Barkow, I. (2017). "Educational process mining: new possibilities for understanding students' problem-solving skills" in *The nature of problem solving. Using research to inspire 21st century learning*. eds. B. Csapó, and J. Funke (Paris: OECD), 193–209).

Vainikainen, M. P., Hautamäki, J., Hotulainen, R., and Kupiainen, S. (2015). General and specific thinking skills and schooling: preparing the mind to new learning. *Think. Skills Creat.* 18, 53–64. doi: 10.1016/j.tsc.2015.04.006

Vígh, T., Sominé Hrebik, O., Thékes, I., and Vidákovich, T. (2015). "Fiatal nyelvtanulók német és angol alapszókincsének diagnosztikus vizsgálata [Diagnostic assessment of young learners' basic English and German vocabulary]" in *Online diagnosztikus mérések az iskola kezdő szakaszában [Online diagnostic assessments in the beginning phase of schooling]*. eds. B. Csapó, and A. Zsolnai (Budapest: Oktatáskutató és Fejlesztő Intézet), 13–33.

Westbury, I., Hansén, S. E., Kansanen, P., and Björkvist, O. (2005). Teacher education for research-based practice in expanded roles: Finland's experience. *Scand. J. Educ. Res.* 49, 475–485. doi: 10.1080/00313830500267937

Wilson, M., Bejar, I., Scalise, K., Templin, J., Wiliam, D., and Irribarra, T. (2012). "Perspectives on methodological issues" in *Assessment and teaching of 21st century skills*. eds. P. Griffin, B. McGaw, and E. Care (New York: Springer), 67–141.

Wu, H., and Molnár, G. (2018a). Interactive problem solving: assessment and relations to combinatorial and inductive reasoning. *J. Psychol. Educ. Res.* 26, 90–105.

Wu, H., and Molnár, G. (2018b). Computer-based assessment of Chinese students' component skills of problem solving: a pilot study. *Int. J. Inf. Educ. Technol.* 8, 381–356. doi: 10.18178/ijiet.2018.8.5.1067

Zsolnai, A., and Kasik, L. (2015). "Az együttműködő viselkedés és az alapérzelem-felismerés online vizsgálata [Online assessment of the recognition of cooperative behavior and basic emotions]" in *Online diagnosztikus mérések az iskola kezdő szakaszában [Online diagnostic assessments in the beginning phase of schooling]*. eds. B. Csapó, and A. Zsolnai (Budapest: Oktatáskutató és Fejlesztő Intézet), 71–95.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Education

TECHNOLOGY REPORT
published: 18 July 2019
doi: 10.3389/feduc.2019.00071

The Argument for a "Data Cube" for Large-Scale Psychometric Data

Alina A. von Davier, Pak Chung Wong, Steve Polyak and Michael Yudelson*

ACTNext, ACT Inc, Iowa City, IA, United States

In recent years, work with educational testing data has changed due to the affordances provided by technology, the availability of large data sets, and by the advances made in data mining and machine learning. Consequently, data analysis has moved from traditional psychometrics to computational psychometrics. Despite advances in the methodology and the availability of the large data sets collected at each administration, the way assessment data is collected, stored, and analyzed by testing organizations is not conducive to these real-time, data intensive computational methods that can reveal new patterns and information about students. In this paper, we propose a new way to label, collect, and store data from large scale educational learning and assessment systems (LAS) using the concept of the "data cube." This paradigm will make the application of machine-learning, learning analytics, and complex analyses possible. It will also allow for storing the content for tests (items) and instruction (videos, simulations, items with scaffolds) as data, which opens up new avenues for personalized learning. This data paradigm will allow us to innovate at a scale far beyond the hypothesis-driven, small-scale research that has characterized educational research in the past.

Keywords: database alignment, learning analytics, diagnostic models, learning pathways, data standards

OPEN ACCESS

Edited by:
Frank Goldhammer,
German Institute for International Educational Research (LG), Germany

Reviewed by:
Pei Sun,
Tsinghua University, China
Hendrik Drachsler,
German Institute for International Educational Research (LG), Germany

****Correspondence:***
Alina A. von Davier
Alina.vonDavier@act.org

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Education

Received: *19 November 2018*
Accepted: *03 July 2019*
Published: *18 July 2019*

Citation:
von Davier AA, Wong PC, Polyak S and Yudelson M (2019) The Argument for a "Data Cube" for Large-Scale Psychometric Data. Front. Educ. 4:71. doi: 10.3389/feduc.2019.00071

INTRODUCTION

In recent years, work with educational testing data has changed due to the affordances provided by technology, availability of large data sets, and due to advances made in data mining and machine learning. Consequently, data analysis moved from traditional psychometrics to computational psychometrics. In the computational psychometrics framework, psychometric theory is blended with large scale, data-driven knowledge discovery (von Davier, 2017). Despite advances in the methodology and the availability of the large data sets collected at each test administration, the way the data (from multiple test forms at multiple test administrations) is currently collected, stored and analyzed by testing organizations is not conducive to these real-time, data intensive computational psychometrics and analytics methods that can reveal new patterns and information about students.

In this paper we primarily focus on data collected from large-scale standardized testing programs that have been around for decades and that have multiple administrations per year. Recently, many testing organizations have started to consider including performance or activity-based tasks in the assessments, developing formative assessments, or embedding assessments into the learning process, which led to new challenges around the data governance: data design, collection, alignment, and storage. Some of these challenges have similarities with those encountered and addressed in the field of learning analytics, in which multiple types of data are merged to provide a comprehensive picture of students' progress. For example, Bakharia et al. (2016), Cooper (2014) and Rayon et al. (2014) propose solutions for the interoperability of learning data coming from multiple sources. In recent years, the testing organizations started to work with logfiles and even before the data exchange standards for activities and events, such as the Caliper or xAPI standards, have been developed, researchers have worked on designing the data schema for this type of rich data (see Hao et al., 2016). The approach presented in this paper conceptually builds on these approaches, being focused on the data governance for testing organizations.

Database Alignment

In this paper, we propose a new way to label, collect, and store data from large scale educational learning and assessment systems (LAS) using the concept of the "data cube," which was introduced by data scientists in the past decade to deal with big data stratification problems in marketing contexts. This concept is also mentioned by Cooper (2014) in the context of interoperability for learning analytics. In statistics and data science the data cube is related to the concept of database alignment, where multiple databases are aligned on various dimensions under some prerequisites (see Gilbert et al., 2017). Applying this paradigm to educational test data is quite challenging, due to the lack of coherence of traditional content tagging, of a common identity management system for test-takers across testing instruments, of collaboration between psychometricians and data scientists, and until recently, of the lack of proven validity of the newly proposed machine learning methods for measurement. Currently, data for psychometrics is stored and analyzed as a two-dimensional matrix—item by examinee. In the time of big data, the expectation is not only that one has access to large volumes of data, but also that the data can be aligned and analyzed on different dimensions in real time—including various item features like content standards.

The best part is that the testing data available from the large testing organizations is valid (the test scores measure what they are supposed to measure, and these validity indices are known) and data privacy policies have been followed appropriately when the data was collected. These are two important features that support quality data and the statistical alignment of separate databases (see Gilbert et al., 2017).

Data Cubes

The idea of relational databases has evolved over time, but the paradigm of the "data cube" is easy to describe. Obviously, the "data cube" is not a cube, given that different data-vectors are of different lengths. A (multidimensional) data cube is designed to organize the data by grouping it into different dimensions, indexing the data, and precomputing queries frequently. Psychometricians and data scientists can interactively navigate their data and visualize the results through slicing, dicing, drilling, rolling, and pivoting, which are various ways to query the data in a data science vocabulary. Because all the data are indexed and precomputed, a data cube query often runs significantly faster than standard queries. Once a data cube is built and precomputed, intuitive data projections on different dimensions can be applied to it through a number of operations. Traditional psychometric models can also be applied at scale and in real time in ways which were not possible before.

Content as Data

Additionally, in this paper we expand the traditional definition of educational data (learning and testing data) to include the content (items, passages, scaffolding to support learning), taxonomies (educational standards, domain specification), the items' metadata (including item statistics, skills and attributes associated with each item), alongside the students' demographics, responses, and process data. Rayon et al. (2014) and Bakharia et al. (2016) also proposed including the content and context for learning data in their data interoperability structures for learning analytics, Scalable Competence Assessment through a Learning Analytics approach (SCALA), and Connected Learning Analytics (CLA) tool kit, respectively. The difference from their approach is in the specifics of the content for tests (items), usage in psychometrics (item banks with metadata), and domain structures such as taxonomies or learning progressions. In addition, we propose a natural language processing (NLP) perspective on these data types that facilitates the analysis and integration with the other types of data.

Any meaningful learning and assessment system is based on a good match of the samples of items and test takers, in terms of the difficulty and content on the items' side, and ability and educational needs on the students' side. In order to facilitate this match at scale, the responses to the test items, the items themselves and their metadata, and demographic data, need to be aligned. Traditionally, in testing data, we collected and stored the students' responses and the demographic data, but the items, instructional content, and the standards have been stored often as a narrative and often it has not been developed, tagged, or stored in a consistent way. There are numerous systems for authoring test content, from paper-based, to Excel spreadsheets, to sophisticated systems. Similarly, the taxonomies or theoretical frameworks by which the content is tagged are also stored in different formats and systems, again from paper to open-sources systems, such as OpenSALT. OpenSALT is an Open source **S**tandards **AL**ignment **T**ool that can be used to inspect, ingest, edit, export and build crosswalks of standards expressed using the IMS Global Competencies and Academic Standards Exchange (CASE) format; we will refer to data standards and models in more detail later in the paper. Some testing programs have well-designed item banks where the items and their metadata are stored, but often the content metadata is not necessarily attached to a taxonomy.

We propose that we rewrite the taxonomies and standards as data in NLP structures that may take the form of sets, or mathematical vectors, and add these vectors as dimensions to the "data cube." Similarly, we should vectorize the items' metadata and/or item models and align them on different dimensions of the "cube."

Data Lakes

The proposed data cube concept could be embedded within the larger context of psychometric data, such as ACT's data lake. At ACT, we are building the **LE**arning **A**nalytics **P**latform (LEAP) for which we proposed an updated version of this data-structure: the in-memory database technology that allows for newer interactive visualization tools to query a higher number of data dimensions interactively. A data lake is a storage solution based on an ability to host large amounts of unprocessed, raw data in the format the sender provides. This includes a range of data representations such as structured, semi-structured, and unstructured. Typically, in a data lake solution, the data structure, and the process for formally accessing it, are not defined until the point where access is required. An architecture for a data lake is typically based on a highly distributed, flexible, scalable storage solution like the Hadoop Distributed File System (HDFS). These types of tools are becoming familiar to testing organizations, as the volume and richness of event data increase. They also facilitate a parallel computational approach for the parameter estimation of complex psychometric models applied to large data sets (see von Davier, 2016).

Data Standards for Exchange

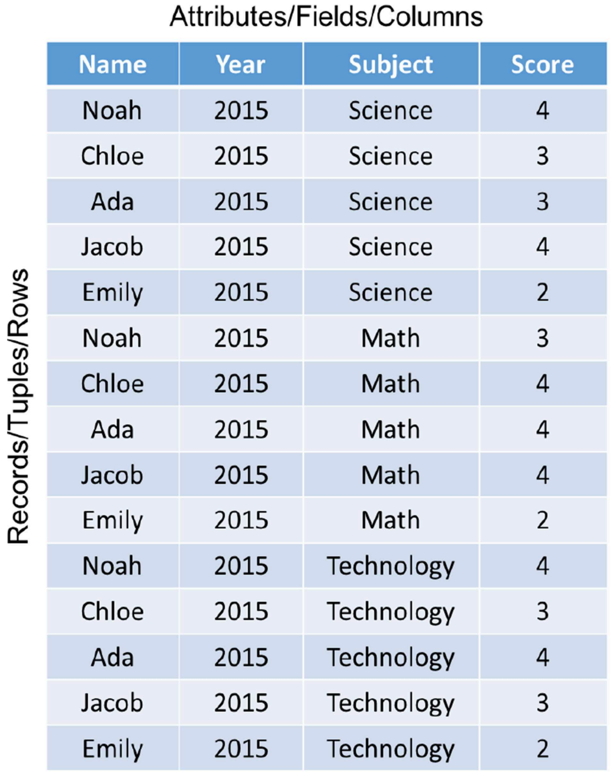
Data standards allow those interoperating in a data ecosystem to access and work with this complex, high-dimensional data (see for example, Cooper, 2014). There are several data standards that exist in the education space which allow schools, testing, and learning companies to share information and build new knowledge, such as combining the test scores with the GPA, attendance data, and demographics for each student in order to identify meaningful patterns that may lead to differentiated instructions or interventions to help students improve. We will describe several of these standards and emphasize the need for universal adoption of data standards for better collaboration and better learning analytics at scale.

In the rest of the paper, we describe the evolution of data storage and the usefulness of the data cube paradigm for large-scale psychometric data. We then describe the approach we are considering for testing and learning data (including the content). In the last section, we present preliminary results from a real-data example of the alignment of two taxonomies from the taxonomy-dimension in the "data cube."

THE FOUNDATIONS OF THE DATA CUBE AND ITS EXTENSIONS

Background and Terminology

In computer science literature, a data cube is a multi-dimensional data structure, or a data array in a computer programming context. Despite the implicit 3D structural concept derived from the word "cube," a data cube can represent any number of data dimensions such as 1D, 2D… nD. In scientific computing studies, such as computational fluid dynamics, data structures similar to a data cube are often referred to as scalars (1D), vectors (2D), or tensors (3D). We will briefly discuss the concept of the relational data model (Codd, 1970) and the corresponding relational databases management system (RDBMS) developed in the 70's, followed by the concept of the data warehouse (Inmon, 1992; Devlin, 1996) developed in the 80's. Together they contributed to the development of the data cube (Gray et al., 1996) concept in the 90's.


Attributes/Fields/Columns
Records/Tuples/Rows

Name	Year	Subject	Score
Noah	2015	Science	4
Chloe	2015	Science	3
Ada	2015	Science	3
Jacob	2015	Science	4
Emily	2015	Science	2
Noah	2015	Math	3
Chloe	2015	Math	4
Ada	2015	Math	4
Jacob	2015	Math	4
Emily	2015	Math	2
Noah	2015	Technology	4
Chloe	2015	Technology	3
Ada	2015	Technology	4
Jacob	2015	Technology	3
Emily	2015	Technology	2

FIGURE 1 | A relational database.

Relational Data Model and Relational Databases Management System (RDBMS)

In a relational data model, data are stored in a table with rows and columns that look similar to a spreadsheet, as shown in **Figure 1**. The columns are referred to as attributes or fields, the rows are called tuples or records, and the table that comprises a set of columns and rows is the relation in RDMBS literature.

The technology was developed when CPU speed was slow, memory was expensive, and disk space was limited. Consequently, design goals were influenced by the need to eliminate the redundancies (or duplicated information), such as "2015" in the Year column in **Figure 1**, through the concept of normalization. The data normalization process involves breaking down a large table into smaller ones through a series of normal forms (or procedures). The discussion of the normalization process is important, but beyond the scope of this paper. Readers are referred to Codd (1970) for further details.

Information retrieval from these normalized tables can be done by joining these tables through the use of unique keys identified during the normalization process. The standard RDBMS language for maintaining and querying a relational database is Structured Query Language (SQL). Variants of SQL can still be found in most modern day databases and spreadsheet systems.

Data Warehousing

The concept of data warehousing was presented by Devlin and Murphy in 1988, as described by Hayes (2002). A data warehouse is primarily a data repository from one or more disparate sources, such as marketing or sales data. Within an enterprise system, such as those commonly found in many large organizations, it is not uncommon to find multiple systems operating independently, even though they all share the same stored data for market research, data mining, and decision support. The role of data warehousing is to eliminate the duplicated efforts in each decision support system. A data warehouse typically includes some business intelligence tools, tools to extract, transform, and load data into the repository, as well as tools to manage and retrieve the data. Running complex SQL queries on a large data warehouse, however, can be time consuming and too costly to be practical.

Data Cube

Due to the limitations of the data warehousing described above, data scientists developed the data cube. A data cube is designed to organize the data by grouping it into different dimensions, indexing the data, and precomputing queries frequently. Because all the data are indexed and precomputed, a data cube query often runs significantly faster than a standard SQL query. In business intelligence applications, the data cube concept is often referred to as Online Analytical Processing (OLAP).

Online Analytical Processing (OLAP) and Business Intelligence

The business sector developed OnLine Analytical Processing technology (OLAP) to conduct business intelligence analysis and look for insights. An OLAP data cube is indeed a multidimensional array of data. For example, the data cube in **Figure 2** represents the same relational data table shown in **Figure 1** with scores from multiple years (i.e., 2015–2017) of the same five students (Noah, Chloe, Ada, Jacob, and Emily) in three academic fields (Science, Math, and Technology). Once again, there is no limitation on the number of dimensions within an OLAP data cube; the 3D cube in **Figure 2** is simply for illustrative purposes. Once a data cube is built and precomputed, intuitive data projections (i.e., mapping of a set into a subset) can be applied to it through a number of operations.

Describing data as a cube has a lot of advantages when analyzing the data. Users can interactively navigate their data and visualize the results through slicing, dicing, drilling, rolling, and pivoting.

Slicing

Given a data cube, such as the one shown in **Figure 2**, users can, for example, extract a part of the data by slicing a rectangular portion of it from the cube, as highlighted in blue in **Figure 3A**. The result is a smaller cube that contains only the 2015 data in **Figure 3B**. Users can slice a cube along any dimension. For example, **Figure 4** shows an example of slicing along the Name dimension highlighted in blue, and **Figure 5** shows an example of slicing along the Subject dimension.

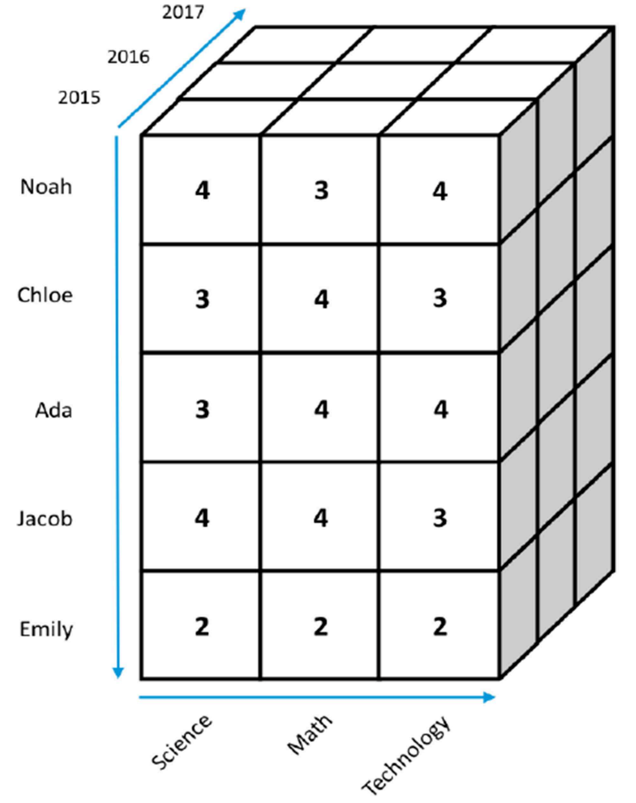


FIGURE 2 | A 3D data cube.

Dicing

The dicing operation is similar to slicing, except dicing allows users to pick specific values along multiple dimensions. In **Figure 6**, the dicing operation is applied to both Name (Chloe, Ada, and Jacob) and Subject (Calculus and Algebra) dimensions. The result is a small 2 × 3 × 3 cube shown in the second part of **Figure 6**.

Drilling

Drilling-up and -down are standard data navigation approaches for multi-dimensional data mining. Drilling-up often involves an aggregation (such as averaging) of a set of attributes, whereas drilling-down brings back the details of a prior drilling-up process.

The drilling operation is particularly useful when dealing with core academic skills that can be best described as a hierarchy. For example, **Figure 7A** shows four skills of Mathematics (i.e., *Number and Quantity*; *Operations, Algebra, and Functions*; *Geometry and Measurement*; and *Statistics and Probability*) as defined by the ACT Holistic Framework (Camara et al., 2015). Each of these skill sets can be further divided into finer sub-skills. **Figure 7B** shows an example of dividing the *Number and Quantity* skill from **Figure 7A** into eight sub-skills—from *Counting and Cardinality* to *Vectors and Matrices*.

Figure 8 shows a drill-down operation in a data cube that first slices along the Subject dimension with the value "Math."

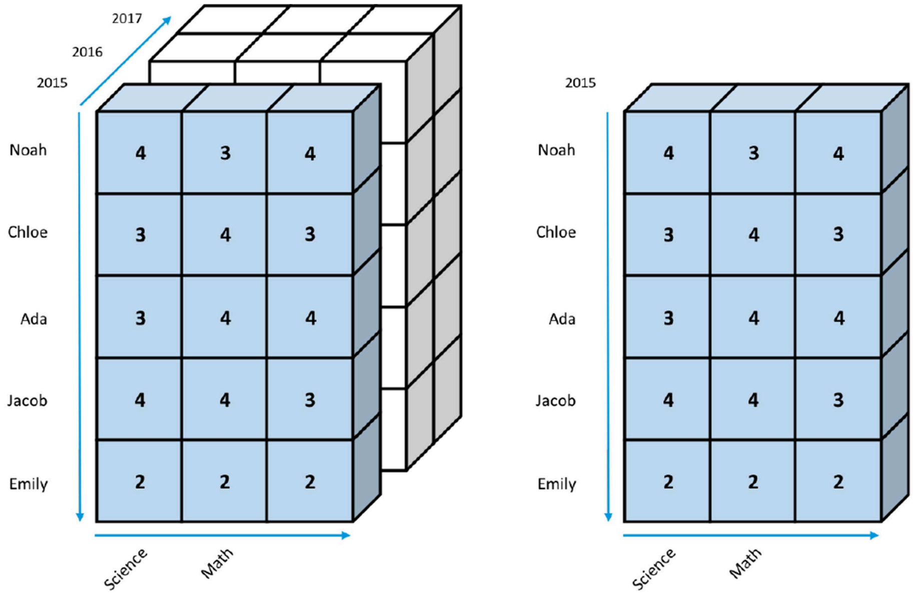


FIGURE 3 | (A,B) Slicing along the Year dimension of a data cube.

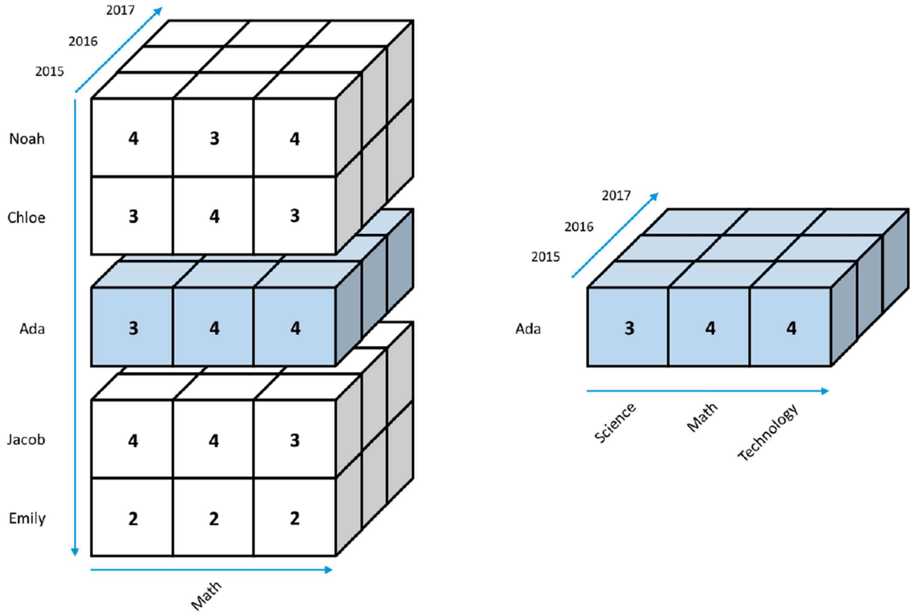


FIGURE 4 | Slicing along the Name dimension of a data cube.

The result is a slice of only the Math scores for all five names from 2015 to 2017 in **Figure 8**. The drilling-down operation in **Figure 8** then shows the single Math score that summarizes the three different Math sub-scores of Calculus, Algebra, and Topology. For example, Emily's 2015 Math score is 2, which is an average of his Calculus (1), Algebra (3), and Topology (2) scores as depicted in **Figure 8**.

The drilling-up operation can go beyond aggregation and can apply rules or mathematical equations to multiple dimensions of a cube and create a new dimension for the cube. The idea, which is similar to the application of a "function" on a spreadsheet, is often referred to as "rolling-up" a data cube.

Pivoting

Pivoting a data cube allows users to look at the cube via different perspectives. **Figure 9** depicts an example of pivoting the data cube from showing the Name vs. Subject front view in the first part of **Figure 9** to a Year vs. Subject in the third part of **Figure 9**,

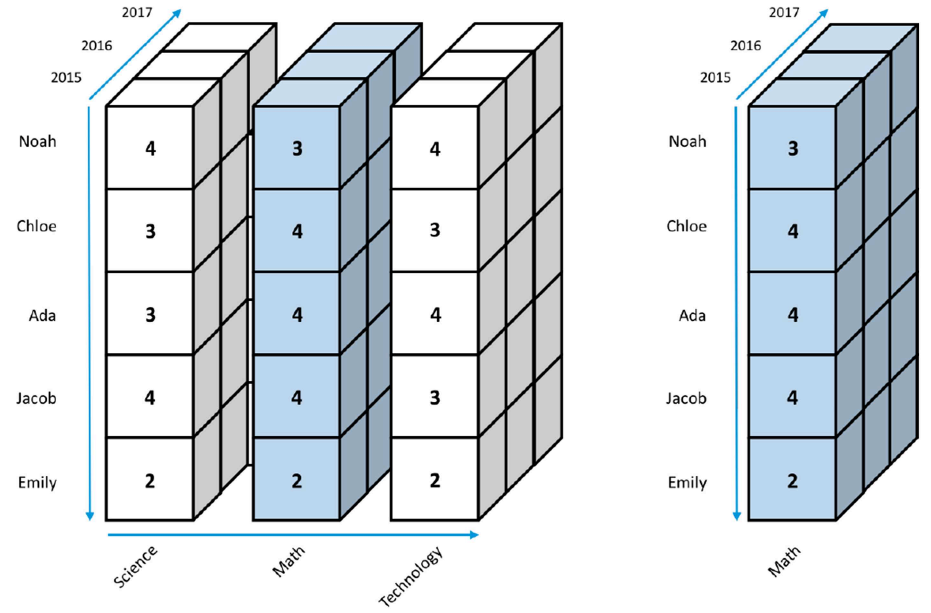


FIGURE 5 | Slicing along the Subject dimension of a data cube.

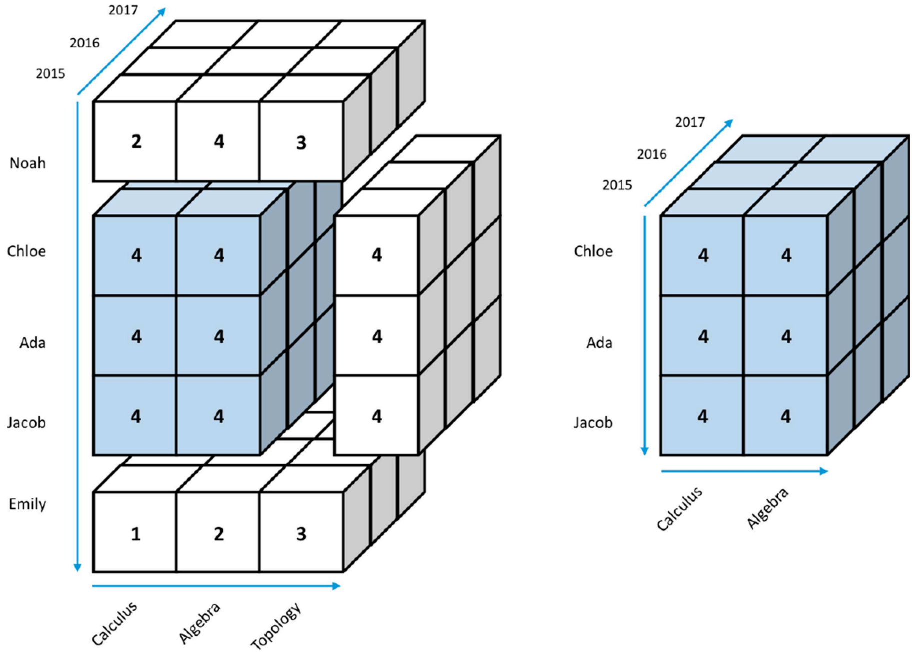


FIGURE 6 | Dicing a 3D data cube.

which shows not just Emily's 2015 scores but also scores from 2016 and 2017. The 3D data cube is indeed rotated backward along the Subject dimension from the middle image to the last image in **Figure 9**.

Beyond Data Cubes

Data cube applications, such as OLAP, take advantage of pre-aggregated data along dimension-levels and provide efficient database querying using languages such as MDX (2016). The more pre-aggregations done on the disk, the better the performance for users. However, all operations are conducted at disk level, which involves slow operation, and thus CPU load and latency issues. As the production cost of computer memory continues to go down and its computational performance continues to go up simultaneously, it has become evident that it is more practical to query data in the

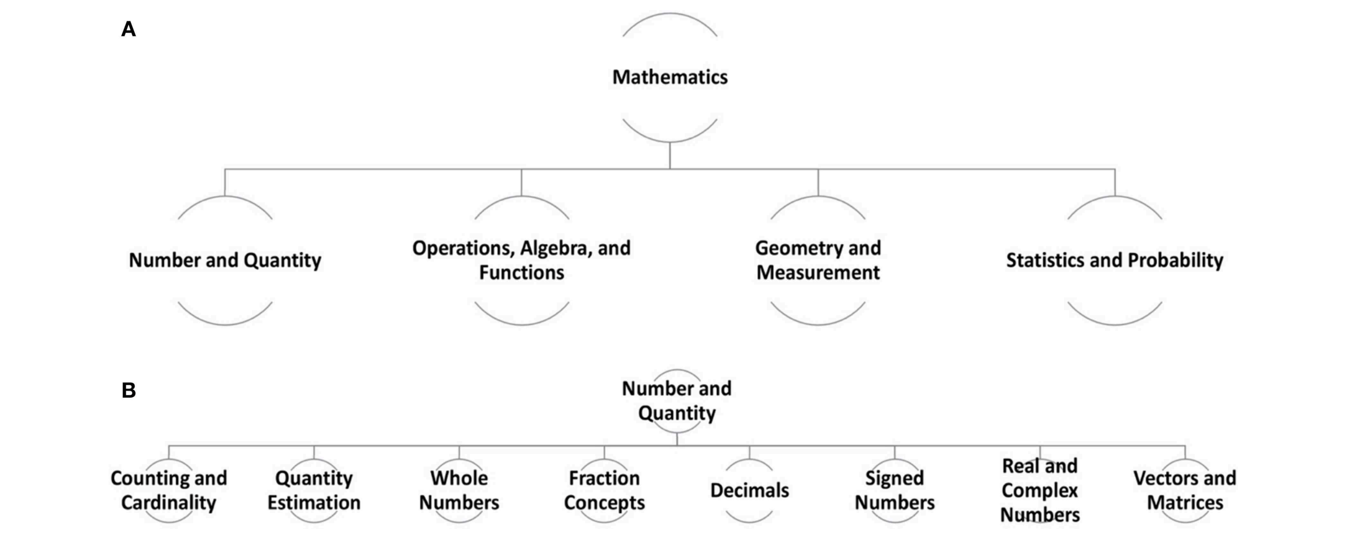


FIGURE 7 | (A) Four skills of Mathematics. **(B)** Eight sub-skills of the Number and Quantity skill.

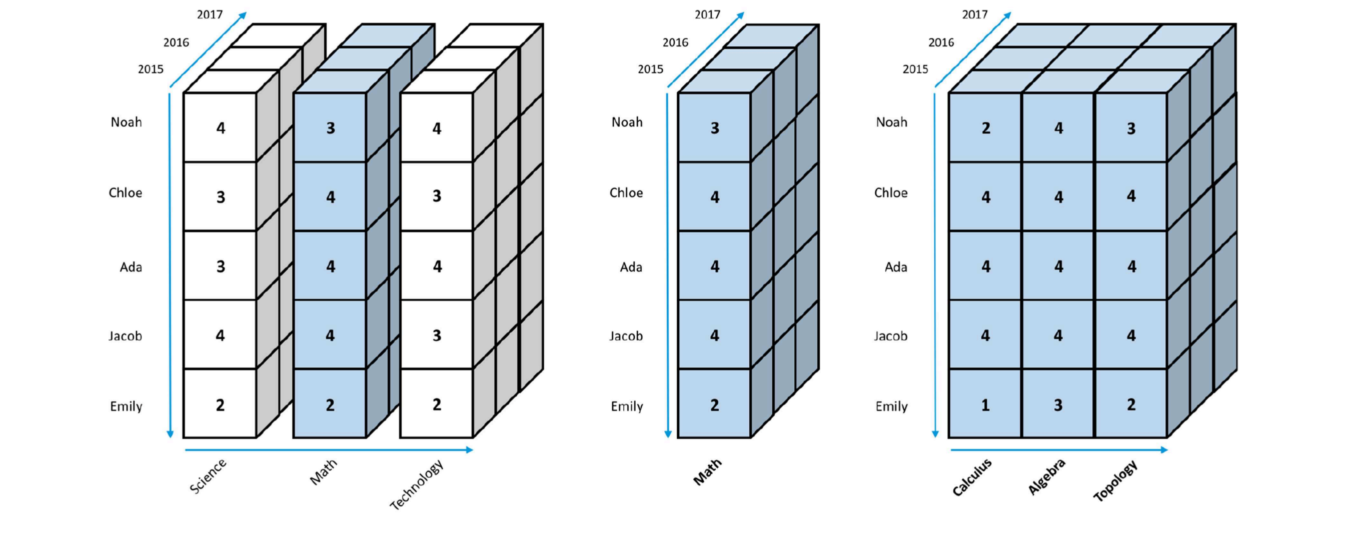


FIGURE 8 | Drilling-down of a data cube.

memory instead of pre-aggregating data on the disk as OLAP data-cubes.

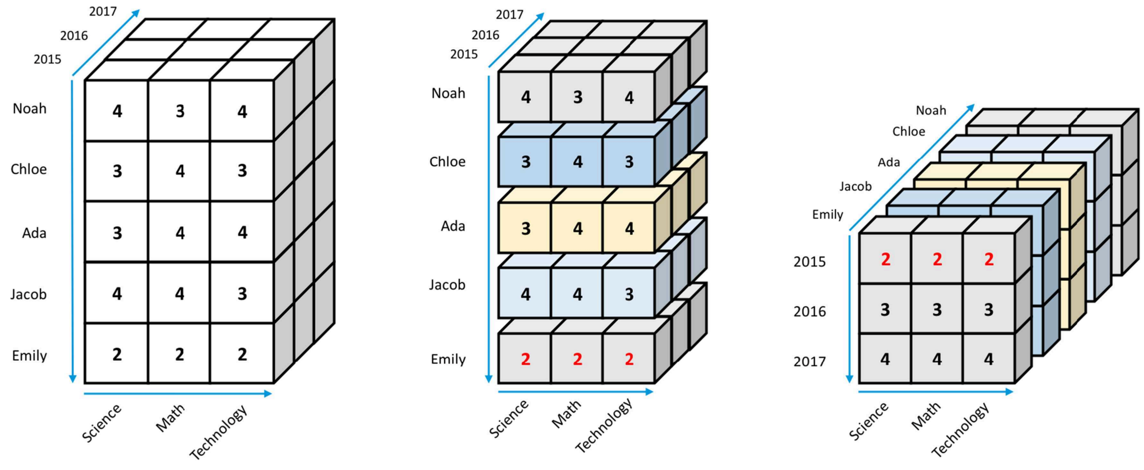
In-memory Computation

Today, researchers use computer clusters with as much as 1 TB of memory (or more) per computer node for high dimensional, in-memory database queries in interactive response time. For example, T-Rex (Wong et al., 2015) is able to query billions of data records in interactive response time using a Resource Description Framework[1] RDF 2014 database and the SPARQL (2008) query language running on a Linux cluster with 32 nodes of Intel Xeon processors and ∼24.5 TB of memory installed across the 32 nodes. Because such a large amount of information can be queued from a database in interactive time, the role of data warehouses continues to diminish in the big data era and as cloud computing becomes the norm.

The Traditional Data Cubes Concept

Additionally, in-memory database technology allows researchers to develop newer interactive visualization tools to query a higher number of data dimensions interactively, which allows users to look at their data simultaneously from different perspectives. For example, T-Rex's "data facets" design, as shown in **Figure 10A**, shows seven data dimensions of a cybersecurity benchmark dataset available in the public domain. After the IP address 172.10.0.6 (in the SIP column) in

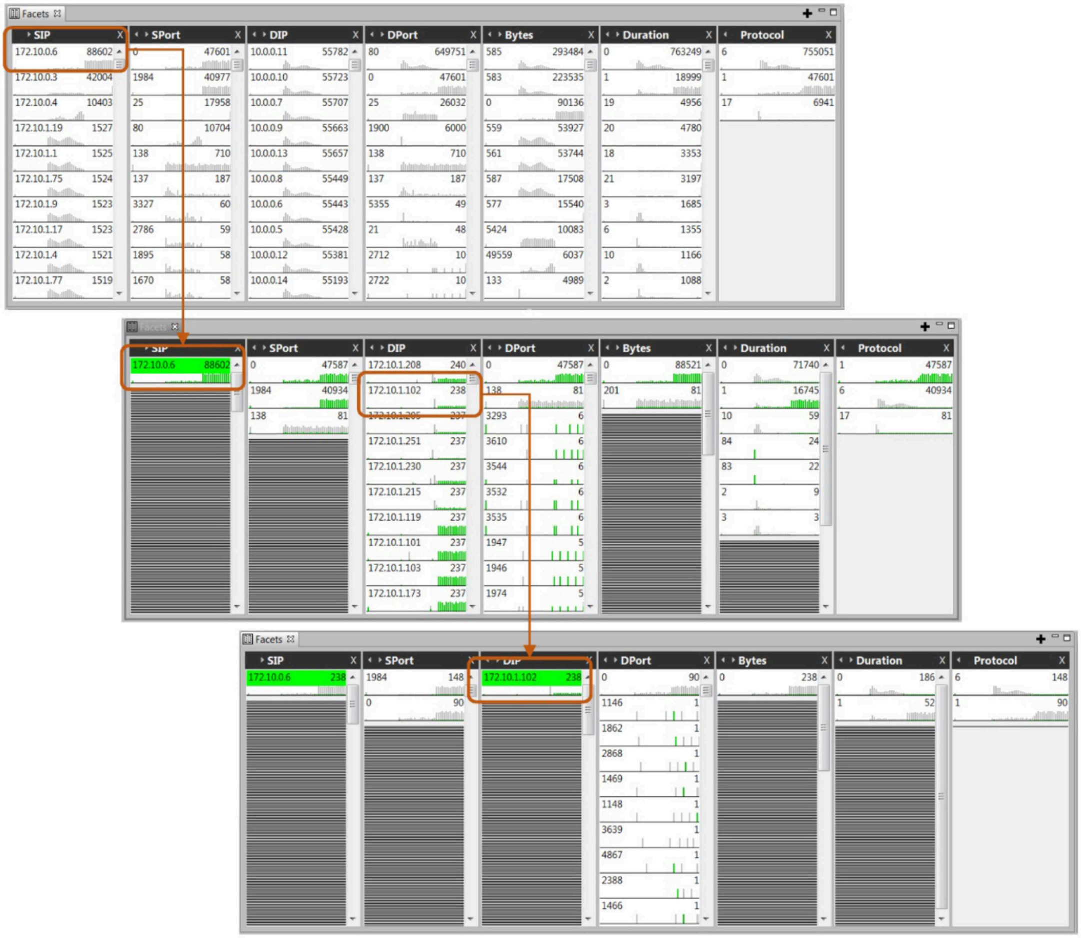
[1] https://en.wikipedia.org/wiki/Resource_Description_Framework



	Science	Math	Technology
Noah	4	3	4
Chloe	3	4	3
Ada	3	4	4
Jacob	4	4	3
Emily	2	2	2

	Science	Math	Technology
2015	2	2	2
2016	3	3	3
2017	4	4	4

FIGURE 9 | Pivoting a data cube from one perspective (dimensional view) to another.



SIP	SPort	DIP	DPort	Bytes	Duration	Protocol
172.10.0.6 88602	0 47601	10.0.0.11 55782	80 649751	585 293484	0 763249	6 755051
172.10.0.3 42004	1984 40977	10.0.0.10 55723	0 47601	583 223535	1 18999	1 47601
172.10.0.4 10403	25 17958	10.0.0.7 55707	25 26032	0 90136	19 4956	17 6941
172.10.1.19 1527	80 10704	10.0.0.9 55663	1900 6000	559 53927	20 4780	
172.10.1.1 1525	138 710	10.0.0.13 55657	138 710	561 53744	18 3353	
172.10.1.75 1524	137 187	10.0.0.8 55449	137 187	587 17508	21 3197	
172.10.1.9 1523	3327 60	10.0.0.6 55443	5355 49	577 15540	3 1685	
172.10.1.17 1523	2786 59	10.0.0.5 55428	21 48	5424 10083	6 1355	
172.10.1.4 1521	1895 58	10.0.0.12 55381	2712 10	49559 6037	10 1166	
172.10.1.77 1519	1670 58	10.0.0.14 55193	2722 10	133 4989	2 1088	

SIP	SPort	DIP	DPort	Bytes	Duration	Protocol
172.10.0.6 88602	0 47587	172.10.1.208 240	0 47587	0 88521	0 71740	1 47587
	1984 40934	172.10.1.102 238	138 81	201 81	1 16745	6 40934
	138 81	172.10.1.205 237	3293 6		10 59	17 81
		172.10.1.251 237	3610 6		84 24	
		172.10.1.230 237	3544 6		83 22	
		172.10.1.215 237	3532 6		2 9	
		172.10.1.119 237	3535 6		3 3	
		172.10.1.101 237	1947 5			
		172.10.1.103 237	1946 5			
		172.10.1.173 237	1974 5			

SIP	SPort	DIP	DPort	Bytes	Duration	Protocol
172.10.0.6 238	1984 148	172.10.1.102 238	0 90	0 238	0 186	6 148
	0 90		1146 1		1 52	1 90
			1862 1			
			2868 1			
			1469 1			
			1148 1			
			3639 1			
			4867 1			
			2388 1			
			1466 1			

FIGURE 10 | Interactive database queries of a high dimensional dataset.

Figure 10A is selected, the data facets update the other six columns as shown in **Figure 10B** simultaneously. The query effort continues in **Figure 10B** where the IP address 172.10.1.102 is queried in the DIP column. **Figure 10C** shows the results after two consecutive queries, shown in green in the figure.

The spreadsheet-like visual layout in **Figure 10** performs more effectively than many traditional OLAP data interfaces found in business intelligence tools. Most importantly, the data facets design allows users to queue data in interactive time without the need for pre-aggregating data with pre-defined options. This video (Pacific Northwest National Laboratory, 2014) shows how T-Rex operates using a number of benchmark datasets available in the public domain.

The general in-memory data cube technology has extensive commercial and public domain support and is here to stay until the next great technology comes along.

DATA CUBE AS PART OF A DATA LAKE SOLUTION AND THE LEAP FOR PSYCHOMETRIC DATA

The proposed data cube concept could be embedded within the larger context of collecting/pooling psychometric data in something that is known in the industry as a data lake (Miloslavskaya and Tolstoy, 2016). An example of this is ACT's data lake solution known as the LEarning Analytics Platform (LEAP). ACT's LEAP is a data lake is a storage solution based on an ability to host large amounts of unprocessed, raw data in the format the sender provides. This includes a range of data representations such as structured, semi-structured, and unstructured. Typically, in a data lake solution, the data structure, and the process for formally accessing it, are not defined until the point where access is required.

A data lake changes the typical process of: extract data, transform it (to a format suitable for querying) and load in to tables (ETL) into one favoring extract, load and transform (ELT), prioritizing the need to capture raw, streaming data prior to prescribing any specific transformation of the data. Thus, data transformation for future use in an analytic procedure is delayed until the need for running this procedure arises. We now describe how the technologies of a data lake help to embed the data cube analysis functionality we described above.

An architecture for a data lake is typically based on a highly distributed, flexible, scalable storage solution like the Hadoop Distributed File System (HDFS). In a nutshell, an HDFS instance is similar to a typical distributed file system, although it provides higher data throughput and access through the use of an implementation of the MapReduce algorithm. MapReduce here refers to the Google algorithm defined in Dean and Ghemawat (2008). ACT's LEAP implementation of this HDFS architecture is based on the industry solution: Hortonworks Data Platform (HDP) which is an easily accessed set of open source technologies. This stores and preserves data in any format given across a set of available servers as data streams (a flow of data) in stream event processors. These stream event processor uses an easy-to-use library for building highly scalable, distributed analyses in real time, such as learning events or (serious) game play events.

Using map/reduce task elements, data scientists and researchers can efficiently handle large volumes of incoming, raw data files. In the MapReduce paradigm:

"Users define the computation in terms of a *map* and a *reduce* function, and the underlying runtime system automatically parallelizes the computation across large-scale clusters of machines, handles machine failures, and schedules inter-machine communication to make efficient use of the network and disk" (Dean and Ghemawat, 2008).

Scripts for slicing, dicing, drilling, and pivoting [See Section Online Analytical Processing (OLAP) and Business Intelligence] in a data cube fashion can be written, executed, and shared via notebook-style interfaces such as those implemented by, for example, open source solutions such as Apache Zeppelin and Jupyter. Zeppelin and Jupyter are web based tools that allow users to create, edit, reuse, and run "data cube"-like analytics using a variety of languages (e.g., R, Python, Scala, etc.). Such scripts can access data on an underlying data source such as HDFS. Organizing analytical code into "notebooks" means combining the descriptive narration of the executed analytical or research methodology along with the code blocks and the results of running them. These scripts are sent to sets of computing machines (called clusters) that manage the process of executing the notebook in a scalable fashion. Data cube applications in the data lake solution typically run as independent sets of processes, coordinated by a main driver program.

Data Standards for Exchange

While data lakes provide flexibility in storage and enable the creation of scaleable data cube analysis, it is also typically a good idea for those operating in a data ecosystem to select a suitable data standard for exchange. This makes it easier for those creating the data, transmitting, and receiving the data to avoid the need to create translations of the data from one system to the next. Data exchange standards allow for the alignment of databases (across various systems), and therefore, facilitate high connectivity of the data stored in the date cube. Specifically, the data exchange standards impose a data schema (names and descriptions of the variables, units, format, etc.) that allow data from multiple sources to be accessed in a similar way.

There are several data standards that exist in the education space that address the data exchange for different types of data, such as:

- Schools Interoperability Framework[2] (SIF) Data Model Specification
- SIF is a data sharing, open specification for academic institutions from kindergarten through workforce. The specification is "composed of two parts: an specification for modeling educational data which is specific to the educational locale, and a system architecture based on both direct and assisted models for sharing that data between institutions, which is international and shared between the locales."
- Ed-Fi Data Standard[3]

  The Ed-Fi Data Standard was developed in order to address the needs of standard integration and organization of data in education. This integration and organization of information

[2] https://en.wikipedia.org/wiki/Schools_Interoperability_Framework (Retrieved May 7, 2018).

[3] https://www.ed-fi.org/

ranges across a broad set of data sources so it can be analyzed, filtered, and put to everyday use in various educational platforms and systems.

- Common Education Data Standards (CEDS)[4]

  CEDS provides a lens for considering and capturing the data standards' relations and applied use in products and services. The area of emphasis for CEDS is on data items and representations across the pre-kindergarten, typical K-12 learning, learning beyond high school, as well as jobs and technical education, ongoing adult-based education, and into workforce areas as well.
- IMS Global[5] Question and Test Interoperability Specification includes many standards. The most popular are the IMS Caliper and CASE.
  - IMS Caliper, which allows us to stream in assessment item responses and processes data that indicate dichotomous outcomes, processes, as well as grade/scoring.
  - IMS Global Competencies and Academic Standards Exchange (CASE), which allows us to import and export machine readable, hierarchical expressions of standards knowledge, skills, abilities and other characteristics (KSAOs). One of the notable examples could be found in (Rayon et al., 2014).
- xAPI – Experience API[6]

  xAPI is a specification for education technology that enables collection of data on the wide range of experiences a person has (both online and offline). xAPI records data in a consistent format about an individual or a group of individual learners interacting with multiple technologies. The vocabulary of the xAPI is simple by design, and the rigor of the systems that are able to securely share data streams is high. On top of regulating data exchange, there exists a body of work toward using xAPI for aligning the isomorphic user data from multiple platforms (rf. Bakharia et al., 2016). An example of aligning activity across multiple social networking platforms is discussed. Also, concrete code and data snippets are given.
- OpenSalt[7]

  We have built and released a tool called OpenSALT which is an Open-source Standards ALignment Tool that can be used to inspect, ingest, edit, export and build crosswalks of standards expressed using the IMS Global CASE format.

  As we outlined in the data cube overview, we are interested in fusing several main data perspectives:
  - Data containing raw item vector analysis data (e.g., correct/incorrect).
  - Data containing complex student-item interactions for item classes beyond assessment.
    - Examples of complex outcomes may include: partial credit results, media interaction results (play), engagement results, and process data (e.g., time spent browsing), tutored interaction, synergetic activities (e.g., interactive labs).
    - Item classes may include: test items, quizzes, and tasks, tutorials, and reading materials.
  - Data that contextualizes this item response analysis within a hierarchical expression of learning objectives/standards collection
    - Item contextualization that addresses multiple hypotheses of how the conceptualization is structured. Multiple hypotheses include accounts for human vs. machine indexing and alternative conceptualizations in the process for development.
  - Demographic data that may include gender, Social and Emotional Skills (SES), locale, and cultural background.
  - Item statistical metadata determined during design and calibration stages (beyond contextualization mentioned above).

The selection of which standards to use to accelerate or enhance the construction of data cubes (within data lakes) for large-scale psychometric data depend on the nature of the educational data for the application. For example, CASE is an emerging standard for injecting knowledge about academic competencies whereas something like xAPI is used to inject the direct feed of learner assessment results (potentially aligned to those CASE-based standards) in a standards-based way into a data cube.

By committing to these data standards, we can leverage the unique capability of the data lake (i.e., efficiently ingesting high volumes of raw data relating to item responses and item metadata) while also prescribing structured commitments to incoming data so that we can build robust, reliable processing scripts. The data cube concept then acts as a high-powered toolset that can take this processed data and enable the online analytical operations such as slicing, dicing, drilling, and pivoting. Moreover, the availability of the data cube and alignment of databases will influence the standards that will need to be available for a smooth integration. It is also possible that new standards will be developed.

EXAMPLE OF APPLICATIONS OF THE DATA CUBE CONCEPT

Alignment of Instruments

One of the key elements of an assessment or learning system is the contextualization of the items and learning activities in terms of descriptive keywords that tie them to the subject. The keywords are often referred to as attributes in the Q-matrices (in psychometrics—see Tatsuoka, 1985), skills, concepts, or tags (in the learning sciences). We will use "concepts" as an overarching term for simplicity. Besides items that psychometrics focuses on, the field of learning sciences has a suite of monikers for elements that cater to learning. The latter include: readings, tutorials, interactive visualizations, and tutored problems (both single-loop and stepped). To cover all classes of deliverable learning

[4] https://en.wikipedia.org/wiki/Common_Education_Data_Standards

[5] https://www.imsglobal.org/aboutims.html

[6] https://xapi.com/overview/

[7] http://opensalt.opened.com/about

and assessment items we would use the term "content-based resources" or "resources" for short.

The relationships between concepts and resources are often referred to as indexing. The intensive labor required to create indexes for a set of items can be leveraged via machine learning/NLP techniques over a tremendous corpus of items/resources. This large scale application was not possible before we had present day storage solutions and sophisticated NLP algorithms. More specifically, the production of said indexing is time-consuming, laborious, and requires trained subject matter experts. There are multiple approaches that address lowering the costs of producing indices that contextualize assessment items and learning resources. These approaches can come in the form a machine learning procedure that, given the training data from an exemplary human indexing, would perform automated indexing of resources.

Data cubes can offer affordances to support the process of production and management of concept-content/resource/item indices. First, even within one subject, such as Math or Science, there could be alternative taxonomies or ontologies that could be used to contextualize resources. See **Figures 7**, **8** for illustrations. Alternatives could come from multiple agencies that develop educational or assessment content or could rely upon an iterative process within one team.

Second, the case when multiple concept taxonomies are used to describe multiple non-overlapping pools of items or resources reserves room for a class of machine learning indexing procedures that could be described as taxonomy alignment procedures. These procedures are tasked with translating between the languages of multiple taxonomies to achieve a ubiquitous indexing of resources.

Third, all classes of machine learning procedures rely upon multiple features within a data cube. The definition and composition of these features is initially developed by subject matter experts. For example, the text that describes the item or resource, its content, or its rationale could be parsed into a high-dimensional linguistic space. Under these circumstances, a deck of binary classifiers (one per concept), or a multi-label classifier could be devised to produce the indexing.

Also, when we are talking about translation form one concept taxonomy to another, one could treat existing expert-produced double-coding of a pool of resources, in terms of the two taxonomies being translated, as a training set. A machine learning procedure, then, would be learning the correspondence relationships. Often, in the form of an *n*-to-*m* mapping example, when one item/resource is assigned *n* concepts from one taxonomy and *m* from the other.

One of our first attempts with translating two alternative concept taxonomies—between the ACT Subject Taxonomy and ACT Holistic Framework—has yielded only modest results. We had only 845 items indexed in both taxonomies and 2,388 items that only had ACT Subject Taxonomy indexing. Active sets of concepts present in the combined set of 3,233 items included 435 and 455 for the Subject Taxonomy and Holistic Framework respectively. A machine learning procedure based on an ensemble of a deck of multinomial regressions (one per each of the 455 predicted Holistic Framework concepts) yielded a 51% adjusted accuracy. Since the index could be sparse, due to the large size of the concept taxonomy and the lower density of items per concept, and the classic machine learning definition of accuracy (matched classifications over total cases classified) would yield an inflated accuracy result due to overwhelming number of cases where the absence of a concept is easily confirmed (we obtained classical accuracies at 99% level consistently). Adjusted accuracy addresses this phenomenon by limiting the denominator to the union of concepts that were present in the human coder-supplied ground-truth training data, or in the prediction (the latter came in the form of pairings of source and target taxonomy concepts, see **Figure 11** for an example). Thus, our work so far and the 51% accuracy should be understood as the first step toward automating taxonomy alignment. We learned that it is significantly harder to align test items than it is to align the instructional resources, because the test items do not usually contain the words that describe the concepts, while the instructional resources do have richer descriptions. This motivated us to include additional data about the test items and the test takers, to increase the samples for the training data, and to refine the models. This is work in progress.

Diagnostic Models

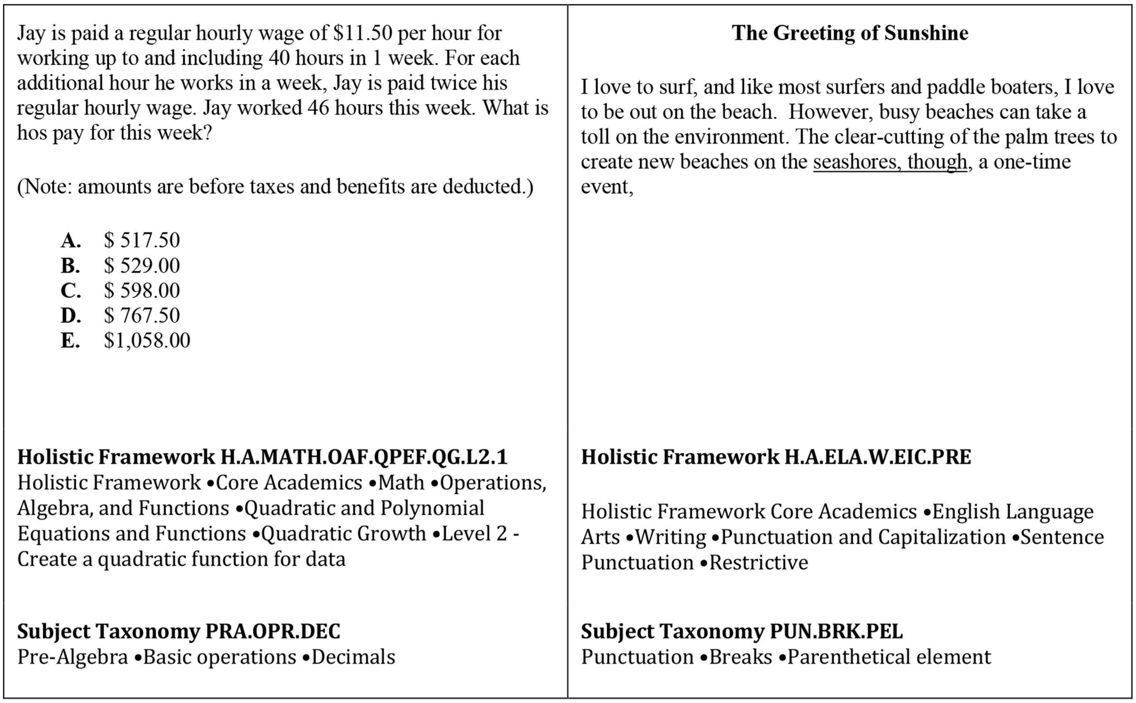
In addition to the alignment of content which is a relatively new application in education, the data cube can support psychometric models that use data from multiple testing administrations and multiple testing instruments. For example, one could develop cognitive diagnostic models (CDMs) that use the data from multiple tests taken by the same individual. CDMs are multivariate latent variable models developed primarily to identify the mastery of skills measured in a particular domain. The CDMs provide fine-grained inferences about the students' mastery and relevance of these inferences to the student learning process.

Basically, a CDM in a data cube relates the response vector $\mathbf{X}_i = \left(X_{i11}, \ldots, X_{ijt}, \ldots, X_{iJ}T\right)$, where X_{ijt} represents the response of the *i*th individual to the *j*th item from the testing instrument *t*, using a lower dimensional discrete latent variable $\mathbf{A}_i = (A_{i1}, \ldots, A_{ik}, \ldots, A_{iK})$ and A_{ik} is a discrete latent variable for individual *i* for latent dimension *k* as described by the taxonomy or the Q-matrix. CDMs model the conditional probability of observing $\mathbf{X}_i$ given $\mathbf{A}_i$, that is, $P(\mathbf{X}_i|\mathbf{A}_i)$. The specific form of the CDM depends on the assumptions we make regarding how the elements of $\mathbf{A}_i$ interact to produce the probabilities of response X_{ijt}.

Traditional data governances in testing organizations cannot easily support the application of the CDMs over many testing administrations and testing instruments: usually the data from each testing instrument is saved in a separate database, that often is not aligned with the data from other instruments. In addition, in the traditional data governance, the taxonomies (and the Q-matrices) across testing instruments are not part of the same framework and are not aligned.

Learning Analytics and Navigation

Another example of the usefulness of a data cube is to provide learning analytics based on the data available about



Jay is paid a regular hourly wage of $11.50 per hour for working up to and including 40 hours in 1 week. For each additional hour he works in a week, Jay is paid twice his regular hourly wage. Jay worked 46 hours this week. What is hos pay for this week? (Note: amounts are before taxes and benefits are deducted.) **A.** $ 517.50 **B.** $ 529.00 **C.** $ 598.00 **D.** $ 767.50 **E.** $1,058.00	**The Greeting of Sunshine** I love to surf, and like most surfers and paddle boaters, I love to be out on the beach. However, busy beaches can take a toll on the environment. The clear-cutting of the palm trees to create new beaches on the seashores, though, a one-time event,
Holistic Framework H.A.MATH.OAF.QPEF.QG.L2.1 Holistic Framework •Core Academics •Math •Operations, Algebra, and Functions •Quadratic and Polynomial Equations and Functions •Quadratic Growth •Level 2 - Create a quadratic function for data	**Holistic Framework H.A.ELA.W.EIC.PRE** Holistic Framework Core Academics •English Language Arts •Writing •Punctuation and Capitalization •Sentence Punctuation •Restrictive
Subject Taxonomy PRA.OPR.DEC Pre-Algebra •Basic operations •Decimals	**Subject Taxonomy PUN.BRK.PEL** Punctuation •Breaks •Parenthetical element

FIGURE 11 | Examples of question items manually tagged with holistic framework and subject taxonomy.

each student. As before, in a data cube, we start with the response vector $\mathbf{X}_i = \left(X_{i11}, \ldots, X_{ijt}, \ldots, X_{iJ}T\right)$, where X_{ijt} represents the response of the ith individual to the jth item from the testing instrument t. Then, let's assume that we also have ancillary data about the student (demographic data, school data, attendance data, etc.) collected in the vector (or matrix) or $\mathbf{B}_i = (B_{i1}, \ldots, B_{im}, \ldots, B_{iM})$ and B_{im} represents a specific type of ancillary variable (gender, school type, attendance data, etc.). Let's assume that for some students we also have data about their success in college, collected under **C**. These data, ***X***, ***B***, and **C** can now be combined across students to first classify all the students, and then later on, to predict the student's success in the first year of college for each student using only the $\mathbf{X}_i$ and $\mathbf{B}_i$. Most importantly, these analytics can be used as the basis for learning pathways for different learning goals and different students to support navigation through educational and career journey.

Learning, Measurement, and Navigation Systems

The ACTNext prototype app, Educational Companion, illustrates an applied instance of linking learning, assessment, and navigation data streams using the data governance described above as the data cube. The app was designed as a mobile solution for flexibly handling the alignment of learner data and content (assessment and instructional) with knowledge and skill taxonomies, while also providing learning analytics feedback and personalized resource recommendations based on the mastery theory of learning to support progress in areas identified as needing intervention. Educational Companion evaluates learning progress by continuously monitoring measurement data drawn from learner interactions across multiple sources, including ACT's portfolio of learning and assessment products. Using test scores from ACT's college readiness exam as a starting point, Companion identifies the underlying relationships between a learner's measurement data and skill taxonomies across core academic areas identified in ACT's Holistic Framework (HF). If available, additional academic assessment data is drawn from a workforce skills assessment (ACT WorkKeys), as well as Socio-Emotional Learning (SEL) data taken from ACT's Tessera exam. Bringing these data streams together, the app predicts skill and knowledge mastery at multiple levels in a taxonomy, such as the HF.

See **Figure 12** for an illustration of the architecture for the Educational Companion App. More details about this prototype are given in von Davier et al. (2019).

As explained in section Alignment of Instruments above, through aligning instructional resources and taxonomic structures using ML and NLP methods, and in conjunction with continuously monitoring updates to a learner's assessment data, Companion uses its knowledge of the learner's predicted abilities along with the understanding of hierarchical, parent/child relationships within the content structure to produce personalized lists of content and drive their learning activities forward. Over time, as learners continue to engage with the app, Companion refines, updates, and adapts its recommendations and predictive analytics to best support an individual learner's needs. The Companion app also incorporates navigational tools developed by Mattern et al. (2017) which

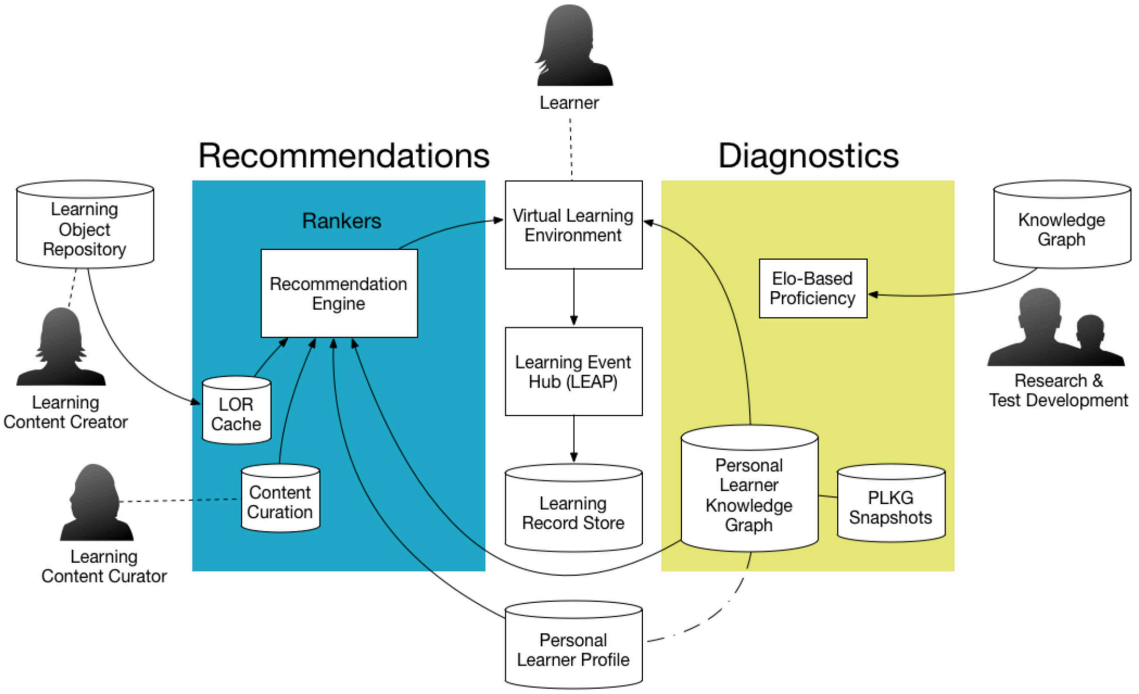


FIGURE 12 | Illustration of the data flow for the ACTNext Educational Companion App. In this figure, the PLKG denotes the personal learning knowledge graph, and the LOR denotes Learning Object Repository. The Elo-based proficiency refers to the estimated proficiency using the Elo ranking algorithm. The knowledge graph is based on the hierarchical relationship of the skills and subskills as described by a taxonomy or standards. A detailed description is available in von Davier et al. (2019).

provide learners with insights related to career interests, as well as the relationships between their personal data (assessment results, g.p.a., etc.) and longitudinal data related to areas of study in college and higher education outcome studies. The Companion app was piloted with a group of Grades 11 and 12 high school students in 2017 (unpublished report, Polyak et al., 2018).

Following the pilot, components from the Educational Companion App were redeployed as capabilities that could extend this methodology to other learning and assessment systems. The ACTNext Recommendation and Diagnostics (RAD) API was released and integrated into ACT's free, online test preparation platform ACT Academy, offering the same mastery theory of learning and free agency via evidence-based diagnostics and personalized recommendations of resources.

CONCLUSION

In this paper we discussed and proposed a new way to structure large-scale psychometric data at testing organizations based on the concepts and tools that exist in other fields, such as marketing and learning analytics. The simplest concept is matching the data across individuals, constructs, and testing instruments in a data cube. We outlined and described the data structure for taxonomies, item metadata, and item responses in this matched multidimensional matrix that will allow for rapid and in-depth visualization and analysis. This new structure will allow real-time, big data analyses, including machine-learning-based alignment of testing instruments, real-time updates of cognitive diagnostic models during the learning process, and real-time feedback and routing to appropriate resources for learners and test takers. The data cube it is almost like Rubik's Cube where one is trying to find the ideal or typical combination of data. There could be clear purposes for that search, for instance creating recommended pathways or recognizing typical patterns for students for specific goals.

In many ways, the large testing companies are well-positioned to create flexible and well-aligned data cubes as described previously. Specifically, the testing data is valid (the test scores measure what they are supposed to measure, and these validity indices are known) and data privacy policies have been followed appropriately when the data was collected, which are two important features that support quality data and the statistical alignment of separate databases. Nevertheless, this new type of data governance has posed challenges for testing organizations. Part of the problem seems to be that the psychometric community has not embraced yet the data governance as part of the psychometrician's duties. The role of this paper is to bring these issues to the attention of psychometricians and underscore the importance of expanding the psychometric tool box to include elements of the data science and governance.

More research and work is needed to refine and improve AI-based methodologies, but without flexible data alignment, the AI-based methods are not possible at all.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

The authors thank Andrew Cantine for his help editing the paper. The authors thank Drs. John Whitmer and Maria Bolsinova for their feedback on the previous version of the paper. The authors thank to the reviewers for their feedback and suggestions.

REFERENCES

Bakharia, A., Kitto, K., Pardo, A., Gašević, D., and Dawson, S. (2016). "Recipe for success: lessons learnt from using xAPI within the connected learning analytics toolkit," in *Proceedings of the Sixth International Conference on Learning Analytics and Knowledge* (ACM), 378–382. doi: 10.1145/2883851.2883882

Camara, W., O'Connor, R., Mattern, K., and Hanson, M.-A. (2015). *Beyond Academics: A Holistic Framework for Enhancing Education and Workplace Success*. ACT Research Report Series (4), ACT, Inc.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Commun. ACM* 13, 377–387. doi: 10.1145/362384.362685

Cooper, A. (2014). *Learning Analytics Interoperability-the Big Picture in Brief*. Learning Analytics Community Exchange.

Dean, J., and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Commun. ACM* 51, 107–113. doi: 10.1145/1327452.1327492

Devlin, B. (1996). *Data Warehouse: From Architecture to Implementation*. Boston, MA: Addison-Wesley Longman Publishing Co., Inc.

Gilbert, R., Lafferty, R., Hagger-Johnson, G., Harron, K., Zhang, L. C., Smith, P., et al. (2017). GUILD: guidance for information about linking data sets. *J. Public Health* 40, 191–198. doi: 10.1093/pubmed/fdx037

Gray, J., Bosworth, A., Layman, A., and Pirahesh, H. (1996). "Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals," in *Proceedings of the International Conference on Data Engineering (ICDE)* (IEEE Computer Society Press), 152–159. doi: 10.1109/ICDE.1996.492099

Hao, J., Smith, L., Mislevy, R., von Davier, A. A., and Bauer, M. (2016). *Taming Log Files From Game/Simulation-Based Assessments: Data Models and Data Analysis Tools*. ETS Research Report Series. Available online at: http://onlinelibrary.wiley.com/doi/10.1002/ets2.12096/full

Hayes, F. (2002). *The Story So Far*. Available online at: https://www.computerworld.com/article/2588199/business-intelligence/the-story-so-far.html

Inmon, W. H. (1992). *Building the Data Warehouse*. New York, NY: John Wiley & Sons, Inc.

Mattern, K., Radunzel, J., Ling, J., Liu, R., Allen, J., and Cruce, T. (2017). *Personalized College Readiness Zone Technical Documentation. Unpublished ACT Technical Manual*. Iowa City, IA: ACT.

MDX (2016). *Multidimensional Expressions (MDX) Reference*. Available online at: https://docs.microsoft.com/en-us/sql/mdx/multidimensional-expressions-mdx-reference

Miloslavskaya, N., and Tolstoy, A. (2016). Big data, fast data and data lake concepts. *Proc. Comput. Sci.* 88, 300–305. doi: 10.1016/j.procs.2016.07.439

Pacific Northwest National Laboratory (2014). *T. Rex Visual Analytics for Transactional Exploration [Video File]*. Retrieved from: https://www.youtube.com/watch?v=GSPkAGREO2E

Polyak, S., Yudelson, M., Peterschmidt, K., von Davier, A. A., and Woo, A. (2018). *ACTNext Educational Companion Pilot Study Report*. Unpublished Manuscript.

Rayon, A., Guenaga, M., and Nunez, A. (2014). "Ensuring the integrity and interoperability of educational usage and social data through Caliper framework to support competency-assessment," in *2014 IEEE Frontiers in Education Conference (FIE) Proceedings* (Madrid: IEEE), 1–9. doi: 10.1109/F.I. E.2014.7044448

RDF (2014). *RDF-Semantic Web Standards*. Available online at: https://www.w3.org/RDF/

SPARQL (2008). *SPARQL Query Language for RDF*. Available online at: www.w3.org/TR/rdf-sparql-query/

Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *J. Educ. Stat.* 12, 55–73. doi: 10.3102/10769986010001055

von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. *J. Educ. Meas.* 54, 3–11. doi: 10.1111/jedm.12129

von Davier, A. A., Deonovic, B., Polyak, S. T., and Woo, A. (2019). Computational psychometrics approach to holistic learning and assessment systems. *Front. Educ.* 4:69. doi: 10.3389/feduc.2019.00069

von Davier, M. (2016). *High-Performance Psychometrics: The Parallel-e Parallel-m Algorithm for Generalized Latent Variable Models*. Princeton, NJ: ETS Research Report. doi: 10.1002/ets2.12120

Wong, P. C., Haglin, D. J., Gillen, D., Chavarria-Miranda, D. G., Giovanni, C., Joslyn, C., et al. (2015). "A visual analytics paradigm enabling trillion-edge graph exploration," in *Proceedings IEEE Symposium on Large Data Analysis and Visualization (LDAV) 2015* (IEEE Computer Society Press), 57–64. doi: 10.1109/LDAV.2015.7348072

Conflict of Interest Statement: AvD, SP, and MY are employed by ACT Inc. PW was employed by ACT Inc. at the time this work was conducted.

frontiers in Psychology

ORIGINAL RESEARCH
published: 23 July 2019
doi: 10.3389/fpsyg.2019.01533

The Onset of Rapid-Guessing Behavior Over the Course of Testing Time: A Matter of Motivation and Cognitive Resources

Marlit Annalena Lindner[1]*, *Oliver Lüdtke*[1,2] *and Gabriel Nagy*[1]

[1] *IPN - Leibniz Institute for Science and Mathematics Education, Kiel, Germany,* [2] *Centre for International Student Assessment, Munich, Germany*

Digital tests make it possible to identify student effort by means of response times, specifically, unrealistically fast responses that are defined as rapid-guessing behavior (RGB). In this study, we used latent class and growth curve models to examine (1) how student characteristics (i.e., gender, school type, general cognitive abilities, and working-memory capacity) are related to the onset point of RGB and its development over the course of a test session (i.e., item positions). Further, we examined (2) the extent to which repeated ratings of task enjoyment (i.e., intercept and slope parameters) are related to the onset and the development of RGB over the course of the test. For this purpose, we analyzed data from $N = 401$ students from fifth and sixth grades in Germany ($n = 247$ academic track; $n = 154$ non-academic track). All participants solved 36 science items under low-stakes conditions and rated their current task enjoyment after each science item, constituting a micro-longitudinal design that allowed students' motivational state to be tracked over the entire test session. In addition, they worked on tests that assessed their general cognitive abilities and working-memory capacity. The results show that students' gender was not significantly related to RGB but that students' school type (which is known to be closely related to academic abilities in the German school system), general cognitive abilities, and their working-memory capacity were significant predictors of an early RGB onset and a stronger RGB increase across testing time. Students' initial rating of task enjoyment was associated with RGB, but only a decline in students' task enjoyment was predictive of earlier RGB onset. Overall, non-academic-school attendance was the most powerful predictor of RGB, together with students' working-memory capacity. The present findings add to the concern that there is an unfortunate relation between students' test-effort investment and their academic and general cognitive abilities. This challenges basic assumptions about motivation-filtering procedures and may threaten a valid interpretation of results from large-scale testing programs that rely on school-type comparisons.

Keywords: rapid-guessing behavior, motivation, test-taking effort, item position effect, low-stakes assessment, large-scale assessment (LSA), latent class analysis

OPEN ACCESS

Edited by:
Ronny Scherer,
University of Oslo, Norway

Reviewed by:
Evangelia Karagiannopoulou,
University of Ioannina, Greece
Grzegorz Szumski,
University of Warsaw, Poland

****Correspondence:***
Marlit Annalena Lindner
mlindner@leibniz-ipn.de

Specialty section:
This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Psychology

Received: *23 November 2018*
Accepted: *18 June 2019*
Published: *23 July 2019*

Citation:
Lindner MA, Lüdtke O and Nagy G
(2019) The Onset of Rapid-Guessing
Behavior Over the Course of Testing
Time: A Matter of Motivation and
Cognitive Resources.
Front. Psychol. 10:1533.
doi: 10.3389/fpsyg.2019.01533

INTRODUCTION

Computer-based assessments are being implemented more and more in educational institutions and large-scale testing programs. This digitalization of tests makes response-time measures (i.e., time on task; e.g., Goldhammer et al., 2014) and log files (e.g., Greiff et al., 2015) easily available. This opens new paths to more objective and also deeper insights into students' test-taking behavior (e.g., Wise and Kong, 2005; Goldhammer et al., 2014; Finn, 2015), for example, by detecting rapid-guessing behavior (RGB). The term RGB basically means that a test-taker provides a response to an item in just a few seconds after the item has been presented. Given that it is highly implausible that students truthfully work on a given task in such a short time frame, RGB is interpreted as a reflection of non-effort (Wise and Kong, 2005; Goldhammer et al., 2016; Wise, 2017). Even though RGB has recently been subject to valuable investigations that shed more light on the nature of this undesirable test-taking behavior, the psychological determinants that are related to RGB in low-stakes assessment have not yet been sufficiently examined.

The present study takes a closer look at the correlates of RGB, placing a specific focus on students' individual probability of showing an early RGB onset over the course of testing time. Specifically, we aimed to investigate the role of two main explanatory psychological characteristics at a student level that are considered to be related to low test-taking effort, namely, a lack of motivational and cognitive resources.

Motivation and Test-Taking Behavior

Educational assessment is essential for the evaluation of learning outcomes and the determination of the proficiency levels of test takers in diverse contexts. Unfortunately, test takers are not always fully motivated to engage in solving test items, especially in low-stakes settings (e.g., Wise and DeMars, 2005, 2010; Wise, 2006; Finn, 2015). Low-stakes means that the test scores have no formal consequences at a student level (e.g., grades, graduation), although aggregated test scores often have major consequences at an institutional or governmental level (e.g., program funding, educational reforms). A high level of effort invested by students when working on a test is considered a prerequisite for a reliable and valid interpretation of achievement levels (Cronbach, 1960; Messick, 1989; Baumert and Demmrich, 2001; Goldhammer et al., 2016). If the problem of low test-taking effort is not treated, for example by statistical correction procedures, students' proficiency may be underestimated, which may lead—in turn—to biased conclusions (see e.g., Wise and DeMars, 2005; Wise et al., 2006b; Nagy et al., 2018b).

Low test-taking motivation in low-stakes assessments is often explained by *Expectancy-Value Models* (e.g., Eccles et al., 1983; Wigfield and Eccles, 2000; Eccles and Wigfield, 2002), which assume that achievement motivation for a given task (e.g., taking a test) is a function of (1) *expectancy* (i.e., students' expectation of success in solving the test items) and (2) *value* (i.e., the perceived importance and usefulness of the test). The expectancy component is determined by both students' abilities and task demands and is, for example, low when test items are too difficult for a student. The value component is considered to be more complex: Eccles and Wigfield (2002) distinguish between four value components, namely, (a) attainment value (e.g., task importance), (b) intrinsic value (e.g., task enjoyment), (c) utility value (e.g., relevance for future goals), and (d) perceived costs (e.g., effort). It can be assumed that all four of these value aspects and, thus, also the overall value component are rather low in low-stakes assessments. This is because, at least for some test takers, the lack of personal consequences and a lack of intrinsic value in taking the test may be in conflict with the effort that is required to successfully solve the items. This is especially true for students with lower competence levels (i.e., low expectancy) who need to invest more effort to successfully work on a test. Accordingly, based on expectancy-value models, achievement motivation can be expected to be lower in low-performing students than in high-performing students.

Lower levels of student motivation become a serious problem when they result in low test effort, which can be defined as a lack of mental work that is put into responding to test items (Wise and DeMars, 2005, 2010; Finn, 2015). Analyzing data sets that include such invalid responses threatens the interpretation of the test scores obtained because construct-irrelevant variance is introduced (Haladyna and Downing, 2004; Nagy et al., 2018a) and psychometric properties are deformed (see e.g., Rios et al., 2017). This issue is often addressed by motivation-filtering procedures (see e.g., Finn, 2015, for a review): As one option, filtering can be based upon self-report questionnaires that aim to assess students' global test-taking motivation (e.g., Student Opinion Scale; Thelk et al., 2009). Such measures are convenient in any type of assessment (including paper-pencil tests), but self-reports are more vulnerable to measurement errors and social desirability (Swerdzewski et al., 2011). As a second option, measuring response times in computer-based assessments provides unobtrusive, more objective insights into students' actual test-taking behavior (e.g., Wise and Kong, 2005; Greiff et al., 2015), while this measure does not disturb or influence students during their taking of the test. Typical sources of measurement error can thus be minimized when referring to students' response behavior as an indicator of effort (or non-effort).

Identifying Rapid-Guessing Behavior

The identification of RGB has proven useful for detecting test takers who do not exert their maximum effort in a test (e.g., Wise, 2006, 2017; Wise et al., 2006b; Finn, 2015). RGB is operationalized by unrealistically low response times that would not even allow the item content to be read and understood and especially would not allow an effortful response; any trial that is not identified as RGB is considered solution behavior, resulting in a dichotomous measure of RGB. However, it is noteworthy that responses that are categorized as solution behavior do not necessarily reflect effortful item solving (for a discussion see e.g., Finn, 2015; Wise, 2017). The main advantage of identifying RGB is that it can be measured for each student and each item. This means that all single trials (i.e., person × item interaction) can be classified as either RGB or solution behavior (see e.g., Wise and Kong, 2005), which makes it possible, for example, to trace the development of non-effort over the course of the test.

However, a reasonable response time threshold needs to be determined to separate (non-effortful) RGB responses from (probably effortful) solution behavior. In doing so, false-positive and false-negative classifications need to be avoided. Various approaches have been discussed (e.g., Wise and Kong, 2005; Wise, 2006; Kong et al., 2007; Wise and Ma, 2012; Lee and Jia, 2014; Finn, 2015; Goldhammer et al., 2016). Defining one constant threshold for every item (e.g., 3 s) is a basic option to determine RGB. However, item-specific, normative thresholds that vary as a function of the mean response time per item (i.e., a certain percentage of the item mean is used to separate RGB from solution behavior; see e.g., Wise and Ma, 2012; Lee and Jia, 2014) or item characteristics (Wise and Kong, 2005; Wise, 2006) often yield a more valid classification of RGB and solution behavior. This is because item attributes can substantially impact the meaning and interpretation of (short) response times. Nonetheless, the different approaches can be helpful in handling different types of data sets (see e.g., Wise, 2017). Thresholds further need to be cross-validated by a combination of different criteria for every test (see e.g., Goldhammer et al., 2016; Wise and Gao, 2017). For example, the accuracy of responses classified as RGB should equal the a priori guessing probability per item, thresholds should be validated by the visual inspection of response time distributions, and 10-s thresholds should not to be exceeded. However, smaller threshold changes do not have a substantial impact on further analyses, suggesting that RGB can be classified with a high reliability—more or less independent of the specific method applied (Kong et al., 2007).

In conclusion, from a pragmatic perspective, RGB can serve as a useful indicator of test-takers' non-effort in motivation-filtering procedures. However, it is also important to gain a better understanding of RGB at a theoretical level and from a psychological point of view.

Theories and Correlates of Rapid-Guessing Behavior

Expectancy-value models help to predict achievement motivation in low-stakes tests. Related assumptions that are more specific to the assessment context and the explanation of RGB have been proposed by Wise and Smith (2011) in the *Demands-Capacity Model* (DCM; see also Wise, 2017). The core of the DCM is the assumption that the tendency of a test taker to engage in RGB is a function of the current fit of (1) the resource demands of the presented items, and (2) the effort capacity of the student. Resource demands are defined as aspects of an item that determine how difficult or mentally taxing it is, such as higher reading demands or complex information. On the other side, test-takers are assumed to have a certain effort capacity that they can invest in solving an item at a specific moment. The DCM is still vague regarding the factors that determine the current status of effort capacity, as the authors propose that many factors have an influence, namely, "test stakes, time pressure, fatigue from answering earlier items, how interesting earlier items were, or a desire to please teachers or parents" (Wise, 2017, p. 53). The DCM further assumes that students compare the current item demands with their current effort capacity. They decide to engage in solution behavior for a given item when their effort capacity is sufficient or, otherwise, to engage in RGB. This explains that test-takers change their response pattern in reaction to different items, as both item demands and effort capacity can easily fluctuate across a test session. Even though RGB is commonly understood as an indicator of a lack of motivation (see e.g., Finn, 2015), building on the DCM, we assume that students might also refuse to work on an item when they lack basic cognitive resources (i.e., as a facet of a lower effort capacity).

Evidence supporting the DCM comes from studies that have investigated correlates of RGB. There are two typical levels of aggregation: the person and the item level. Regarding the student level, the measure of response time effort (RTE[1]), as introduced by Wise and Kong (2005), is determined as the proportion of solution behavior related to all presented items in a test and provides information concerning the overall level of invested effort per student. The correlations of RTE and person characteristics can provide information concerning factors that go along with higher or lower levels of test-taking effort, respectively. The item-specific counterpart, introduced by Wise (2006), is response time fidelity (RTF). It represents the effort invested in a specific item across all test-takers, namely, the proportion of effortful responses to that item. Thus, RTF is a useful parameter to investigate correlates of effort based on item characteristics. It is also possible to model students' responses by more complex linear or generalized mixed-effects models (e.g., Wise et al., 2009) to jointly investigate student and item characteristics and their connections to RGB.

Building on RTE and RTF and using multilevel approaches, research has shown that higher RGB prevalence at a student level (i.e., RTE) is, for example, often associated with lower academic abilities (e.g., Wise et al., 2009; Lee and Jia, 2014; Goldhammer et al., 2016; Wise and Gao, 2017), male gender (e.g., DeMars et al., 2013; Goldhammer et al., 2016), personality traits, such as lower conscientiousness and agreeableness or higher neuroticism (e.g., DeMars et al., 2013; Barry and Finney, 2016; Lu et al., 2018), and cultural background characteristics (e.g., Goldhammer et al., 2016). However, the findings are not consistent across studies. Especially the relation of test effort and academic ability levels needs to be discussed and investigated more as the results are mixed and of high practical importance (see e.g., Wise and DeMars, 2005; Wise and Kong, 2005; Wise et al., 2006b, 2009; Kong et al., 2007; Lee and Jia, 2014; Goldhammer et al., 2016; Wise and Gao, 2017). Overall, previous findings align with the DCM as they suggest that academic and motivational resources as well as sociocultural aspects play a role in test-takers' effort capacity, which is assumed to be responsible for their decisions to show solution behavior or to engage in RGB instead.

Again in line with DCM assumptions, there is evidence that item characteristics (i.e., item demands) influence students' tendency to engage in RGB. Especially surface characteristics, such as shorter texts and the presence of pictures have been shown to be related to lower RGB rates (Wise et al., 2009; Lindner

[1] Wise and Gao (2017) recently proposed a broader measure of test-taking effort, which they refer to as response behavior effort (RBE) and response behavior fidelity (RBF), which makes it possible to identify rapid omits and rapid perfunctory answers to constructed response items in addition to RGB.

et al., 2017a). However, deep item characteristics that are not easily traceable at first sight, such as item difficulty or the content area of an item did not have a significant impact on RGB rates, as shown by Wise et al. (2009). From a logical point of view, this is not surprising because the short time frame in which students look at an item before they engage in RGB is not long enough to analyze deeper item characteristics. Thus, the item appearance seems to be more important for the perception of item demands and the decision to engage in RGB or not.

Furthermore, the circumstances of the test situation have been connected to test-taking effort and RGB rates. For example, although different seasons or weekdays did not influence students' test-taking effort, a later testing time in a day (e.g., testing in the afternoon) was linked to lower RTE measures (i.e., more RGB; Wise et al., 2010). This suggests that physical and/or mental fatigue plays a role in reduced test-taking effort (Lindner et al., 2018), which may also explain why the most important predictor of RGB is the elapsed testing time (see e.g., Wise et al., 2009). There is compelling evidence across studies that items presented in later positions in a test are typically solved with lower accuracy (item position effect; e.g., List et al., 2017; Weirich et al., 2017; Nagy et al., 2018a), less motivational effort (e.g., Barry and Finney, 2016; Penk and Richter, 2017) and are substantially more prone to RGB (e.g., Wise et al., 2009; Setzer et al., 2013; Goldhammer et al., 2016).

Consequentially, because test-item demands change neither with day times nor with the test duration, the existing findings indicate that the reported increase of RGB over the course of testing time is mostly related to changes at the level of test takers' resources. Overall, there is reason to assume that both motivational and cognitive capacities become exhausted over the course of testing time due to the effort that has already been invested in solving previous items. Specifically, students need to build a new situational mental model for every single item and cognitively switch between tasks and solution strategies in a short time frame (Lindner et al., 2017a). Such operations are demanding and require working-memory capacity (i.e., executive attention; Engle, 2002) and self-control (Lindner et al., 2017a). Following Inzlicht et al. (2014), investing self-control to focus attention on cognitive tasks becomes more and more aversive over time, leading to a motivational disengagement from effortful tasks while attentional disruptions increase. This is also presumed to go along with a negative influence on students' affect over the course of a test session, which may cause a reduction in motivational effort (e.g., Ackerman and Kanfer, 2009; Ackerman et al., 2010; Inzlicht et al., 2014). As a consequence, individuals' performance typically decreases over the course of the test (e.g., Penk and Richter, 2017; Nagy et al., 2018b).

In this study, based on the DCM, we assumed that increasing exhaustion and negative emotions would be more pronounced for students who have lower cognitive capacities (i.e., academic abilities, general cognitive abilities, and working-memory capacity) and lower motivational capacities (i.e., low task enjoyment). Thus, we expected that students with lower cognitive and motivational resources suffer from an earlier depletion of their effort capacity and, thus, start to engage in RGB at an earlier point in the testing time.

The Present Research

Although different studies have investigated the correlates of RGB, they mainly considered the frequencies or proportions of RGB (i.e., RTE or RTF; e.g., Wise and Kong, 2005; Wise, 2006) and, to the best of our knowledge, no studies have yet focused on correlates for the RGB onset in a test session. Furthermore, the question remains open of whether a lower level of test motivation at the beginning of the test and a (faster) loss of motivation over the course of the test session are associated with an earlier RGB onset. The present study aimed to answer these questions by investigating the measures of student characteristics (i.e., gender, school type, general cognitive abilities, and working-memory capacity) as well as data from a micro-longitudinal design with 36 repeated ratings of students' task enjoyment over the course of testing. Our main goal was to investigate the relations of these cognitive and motivational measures to students' individual risk of early RGB onset during a test, in order to enhance the theoretical understanding of the RGB phenomenon.

Parts of the underlying data set have been previously published with a much different focus on the effects of representational pictures in testing (see Lindner et al., 2017b). RGB was one of three dependent variables in the investigation of the effects of pictures as an item design characteristic. We do not report the respective findings in this study but, rather, directly build on the prior insights regarding students' RGB development across time, which we summarize here very briefly. In line with the literature (e.g., Goldhammer et al., 2016), the data showed a substantial RGB increase over the course of the test session, indicated by a significant main effect of item position (see Lindner et al., 2017b). However, this increase was substantially smaller in items that contained a representational picture (significant main effect picture). There was no significant interaction between the factors picture and item position. Pictures mainly induced a shift in RGB frequency. Both text-only and text-picture items were subject to an increase in RGB across time, but the probability of RGB was smaller throughout the test for items that contained a picture. In the current analyses, we took the systematic variation of picture presence as a control factor into account, but did not specifically investigate this characteristic.

In line with the literature, we assumed in the present research that RGB is a type of behavior that, similar to other phenomena in the testing context (e.g., item position effects, performance decline; e.g., Hartig and Buchholz, 2012; Debeer et al., 2014; Jin and Wang, 2014; List et al., 2017; Weirich et al., 2017; Wise and Gao, 2017; Nagy et al., 2018b), has a high probability of being maintained (at a student level) over the course of a test session, once it has begun. This means that once individuals engage in RGB, they have a high probability of showing this behavior in the subsequent items of the test. This assumption is also in line with insights from raw data of individuals' RGB development as well as with the DCM (Wise, 2017), according to which a depletion of students' effort capacity across time goes along with a higher probability of engaging in RGB. This hypothesis also formed the base of our attempt to model the data in a latent class approach to investigate the correlates of students' RGB onset, which will be explained in detail in the Methods section. Specifically, drawing on the empirical and theoretical background in the field as outlined above, we formulated the following hypotheses:

> *Hypothesis 1*: We expected to find a higher probability of earlier RGB onset in (a) male students, (b) students from non-academic-track schools, (c) students with lower cognitive resources in terms of general cognitive abilities, and (d) students with lower cognitive resources in terms of working-memory capacity.
>
> *Hypothesis 2:* We expected that both the initial level of students' task enjoyment and its (negative) development over the course of testing would be predictive of RGB. Specifically, we expected that both (a) lower initial enjoyment ratings (intercept) and (b) a stronger decrease (slope) would be associated with the RGB variable and predict earlier RGB onset.

METHODS

As mentioned above, the current data set has been subject to investigations before. To avoid unnecessary repetition, we only report the measures that are relevant for the present analyses. Please consult the report by Lindner et al. (2017b) for further details.

Sample, Material, and Study Design

The analyzed sample comprised $N = 401$ students in the fifth and sixth grades in northern Germany (53.4% female, 51.4% fifth grade, $M_{\text{age}} = 10.74$, $SD_{\text{age}} = 0.76$; $n = 247$ academic track [i.e., Gymnasium]; $n = 154$ non-academic track [i.e., regional school]) who took a computerized science test in an experimental classroom setting. Students were informed that their individual participation was completely voluntary and that they would not face any negative consequences if they did not participate or if they canceled their participation. Thus, all students were fully aware of the low-stakes testing environment, but they were also informed about the relevance of investing good effort to ensure reliable research results.

The scientific literacy test was constructed based on the science framework and items of the Trends in International Mathematics and Science Study (TIMSS; see e.g., Mullis et al., 2009; International Association for the Evaluation of Educational Achievement [IEA], 2013), which assess students' basic science achievement. The 36 items confronted students with realistic situations, forcing them to apply their declarative science knowledge from biology, physics, and chemistry to everyday phenomena and problems. It was essential that the students correctly understood the situation in the item stem for them to be able to solve the problem correctly. The items had a mean word count of $M = 74.9$ words ($SD_{\text{words}} = 24.2$). All items were presented in a multiple-choice format with a short item stem, a separate one-sentence question, and four answer options (one correct option). The items were randomly assigned to one of three test blocks (12 items per block), which were presented either with or without representational pictures (i.e., experimental manipulation of test items), resulting in six booklet constellations. A randomization check confirmed that the item difficulty did not differ between the blocks, $F_{(2,33)} = 0.05$; $p = 0.95$; $\eta^2 = 0.003$. The systematic variation of presenting a representational picture (or not) in the items was balanced across booklets and realized in a within-subject multi-matrix design. To investigate RGB over the course of the test (i.e., in different item positions), items were presented in a random order within test blocks to avoid presenting certain items in certain positions. The six booklets were randomly assigned to the students and equally distributed in the sample (including school types). The marginal EAP/PV reliability of the science test was estimated as $Rel. = 0.83$.

Measures

Background Variables

We used a short questionnaire to assess background information, such as students' age, gender, grade level (fifth vs. sixth grade) and the attended school type (academic and non-academic track).

General Cognitive Abilities

The subtest N2 (Figural Analogies; adjusted according to students' grade level; $\alpha = 0.93/0.89$) of the Kognitiver Fähigkeitstest (KFT) 4 – 12 + R (Heller and Perleth, 2000) was applied to measure spatial reasoning skills as an indicator of students' general cognitive abilities and resources. The subscale consists of 25 items, each of which presents students with one pair of meaningfully related figures and another single figure, for which the appropriate counterpart has to be selected from five answer options in order to create a similar pair of related figures.

Working-Memory Capacity

A self-programmed, computerized version of a reversed digit span test (see e.g., HAWIK-IV; Petermann and Petermann, 2010) served as an indicator of students' working-memory capacity. Students listened through headphones to an increasing number of digits (i.e., two up to eight) that were read out at a slow pace by a male voice. During the digit presentation, the keyboard was locked. After hearing each row (e.g., 3–5–8–7), students were asked to type the digit row in reverse order (e.g., 7,853) into a box that appeared on the screen. After logging in the response, the screen went white and the next digit row followed. The test contained 14 trials. The sum of correct answers determined the test score. Reliability was just sufficient ($\alpha = 0.64$).

Task Enjoyment Ratings

As an indicator of students' current motivational level, we repeatedly measured students' task enjoyment while working on the items. We did so with a one-item measure (see Lindner et al., 2016), asking students how much fun they had solving the current item (i.e., "Working on this item was fun for me"). We assumed that lower enjoyment ratings would indicate lower motivational resources.

Rapid-Guessing Behavior

Students' response time was measured per item (in seconds), which served as the base for classifying RGB trials. Extreme response times two standard deviations (*SD*) above the item mean (0.3% of the data) were trimmed by replacing them with the value of two *SD* above the item mean (e.g., Goldhammer et al., 2014) to prevent bias in the means. Afterwards, the mean time on task for each item served as a base for setting RGB thresholds, following the normative threshold (NT) method proposed by Wise and Ma (2012). Using this method, item-specific threshold percentages can be defined, which means that response times shorter than, for example, 10%, 15%, or 20% of the average solution time of an item are classified as rapid guesses. To achieve a balance between identifying as many non-effortful responses as possible and avoiding the classification of effortful responses as RGB (e.g., Wise and Kong, 2005; Lee and Jia, 2014), we used a mixed approach to evaluate potential thresholds by different validation methods (i.e., absolute thresholds, visual inspection and guessing probability in RGB trials; e.g., Goldhammer et al., 2016; see also section Motivation and Test-Taking Behavior). Taking all validation criteria into account (for a detailed evaluation, see Lindner et al., 2017b), the NT15 criterion turned out to deliver the best fit and was thus used for the RGB definition. This resulted in an average item-specific threshold of $M = 5.6$ s ($SD = 1.4$).

Apparatus and Procedure

Experienced test administrators conducted the study at schools during lesson time. All sessions were attended by a teacher and lasted up to 90 minutes. A laptop, headphones, and a mouse were prepared for each student. The science items were presented on 28 identical Lenovo® laptops, using the software flexSURVEY 2.0 (Hartenstein, 2012). Students answered a short background questionnaire, worked on the KFT, and took the working-memory test. Afterwards, they worked on the science test. It was ensured that students knew that they would not be able to return to an earlier question after choosing an answer and that they always needed to provide a response in order to progress to the next item. Following each item, students rated their item-solving valence. Providing an answer automatically forwarded the student to the next task. Students were repeatedly encouraged to take all the time they needed to solve each item but to work in a focused way through the test. This was done to ensure that the science test was worked on as a power test. There was no time limit for completing the test. Responses, response times (per item), and the item presentation sequence (i.e., item positions) were recorded in a log file for each student.

Data Analyses

RGB is a low-frequency behavior that is not exhibited by each student. As such, statistical modeling approaches for RGB should divide the total sample into at least two groups (or latent classes): One class that does not show RGB at all, and a second class of respondents who show at least some RGB responses. Within the class of individuals showing some RGB, the representation of the distribution of RGB can be challenging, especially in samples of modest size.

As a solution to this problem, we modeled RGB by means of a categorical latent variable (i.e., a latent class analysis; LCA). Our LCA model distinguished between latent classes that showed no RGB at all (i.e., no-RGB class), and three other classes that differed in the onset points of RGB (i.e., early, intermediate, and late onset points). In addition, we assumed the existence of a latent class consisting of students who had a rather low but constant probability of RGB at any point in the test (i.e., constantly low RGB class). To achieve this goal, we modeled the logits of the probability of RGB indicators y_{ip} [$y_{ip} = 1$ if individual i ($i = 1, 2, \ldots, N$) showed RGB in position p ($p = 1, 2, \ldots, 36$), and $y_{ip} = 0$ otherwise] conditional on class membership $C_i = k$ ($k = 1, 2, 3, 4, 5$):

$$\operatorname{logit}\left[P\left(y_{ip} = 1 | C_i = k\right)\right] = \gamma_k w_{ip} + \tau_{0k} + \frac{\theta_k}{1 + \exp\left[-\alpha_k\left(\beta_k - p\right)\right]} \qquad (1)$$

In Equation (1), w_{ip} is a variable indicating whether the item presented in position p to individual i is a text-only ($w_{ip} = 0$) or a text-picture item ($w_{ip} = 1$), and γ_k is a logistic regression weight accounting for the fact that text-picture items are less likely to be associated with RGB (see Wise et al., 2009; Lindner et al., 2017b). The parameter γ_k was specified to be invariant across classes reflecting RGB (i.e., $C = 1$–4), but was constrained to zero in the no-RGB class ($C = 5$). The last two terms on the right-hand side of Equation (1) capture the development of RGB across item positions. τ_{0k} is a lower asymptote parameter, and θ_k describes the upper asymptote of the probability of RGB in class $C = k$. The parameter α_k ($\alpha_k \geq 0$) reflects the rate of change in RGB probabilities, whereas β_k stands for the position in which the inflection point of the logistic function occurs in class $C = k$.

In order to provide an interpretable solution, the LCA parameters of Equation (1) were subjected to further constraints. The first three classes ($C \leq 3$) were specified to reflect students with different onset points of RGB (parameters β_k). Here, we specified the β_k parameters to be ordered (i.e., $\beta_1 < \beta_2 < \beta_3$) and equally spaced, and the lower and upper asymptotes, τ_{0k} and θ_k, to be equal across these three classes. In order to provide an interpretable asymptote parameter, we constrained the rate-of-change parameter α_k in such a way that the RGB probability in $p = 1$ (i.e., first item position) in the late-RGB-onset class ($C = 3$) solely reflected the lower asymptote τ_{0k}. To this end, we constrained the last term of Equation (1) to be very close to zero in $p = 1$ by imposing the constraint $\alpha_k = \frac{logit(0.001)}{(\beta_3 - 1)}$. The constantly low RGB class ($C = 4$) was assumed to have the same τ_{0k} and α_k parameters as the classes $C = 1$ to 3, but θ_4 was allowed to take a different value. In this class, β_4 was set to be equal to the inflection point of the early-RGB-onset class ($C = 1$), β_1. Finally, in the no-RGB class ($C = 5$), the parameters γ_5, θ_5, α_5, and β_5 were fixed to zero, and τ_{05} was fixed to −15. Taken together, our basic LCA model estimated only six measurement parameters (Equation 1), and four latent class proportions π_1 to π_4 ($\pi_5 = 1 - \sum_{k=1}^{K-1} \pi_k$).

The LCA model was extended by the inclusion of covariates predicting class membership. This was accomplished by means of a multinomial logit model so that:

$$P\left(C_i = k | x_i\right) = \frac{\exp\left(\omega_{0k} + \sum_{j=1}^{J} \omega_{1kj} x_{ij}\right)}{\sum_{l=1}^{l=5} \exp\left(\omega_{0l} + \sum_{j=1}^{J} \omega_{1lj} x_{ij}\right)}, \qquad (2)$$

with x_i being the individual i's $J \times 1$ vector of covariate values with entries x_{ij} for covariates $j = 1, 2, \ldots, J$, and ω parameters standing for multinomial intercepts and weights that were fixed to zero for the no-RGB class $C = 5$. Based on the estimates of the ω-parmameters, RGB probability curves, expected at specific values of the covariate x_i, were derived by combining Equations (1, 2) to:

$$P\left(y_{ip} = 1|x_i\right) = \sum_{k=1}^{K} P\left(C_i = k|x_i\right) P\left(y_{ip} = 1|C_i = k\right). \quad (3)$$

Most covariates were observed but, in the case of task enjoyment, we used latent variables that were derived from a linear growth model specified as:

$$z_{ip} = \delta w_{ip} + \eta_{0i} + \frac{p-1}{36-1}\eta_{1i} + \varepsilon_{ip}, \quad (4)$$

where z_{ip} is the individual i's enjoyment score in position p, w_{ip} stands for the values of the item-level covariate as defined before, and δ is a corresponding regression weight. The latent variables η_{0i} and η_{1i} represent the individual's initial enjoyment value and the rate of change, while ε_{ip} is a random disturbance. The η-variables were assumed to follow a bivariate normal distribution. Disturbances were assumed to have zero means, to be normally distributed, and to be uncorrelated with each other as well as with any other variable in the system. The variances of disturbances were set to be equal across positions, but were allowed to be different for text-only and text-picture items. The η-variables were entered into the LCA models similar to x-variables (Equation 2), where all growth and LCA parameters were jointly estimated.

All estimations were carried out with the M*plus* 8.0 program (Muthén and Muthén, 2017) using marginal maximum likelihood estimation. Parameter estimates were accompanied by robust standard errors adjusted for non-normality. As LCA models are known to be prone to local minima, we used multiple random start values to check whether the best log-likelihood could be replicated. Model-data fit was evaluated by information theoretic indices including the Akaike information criterion (AIC), Bayesian information criterion (BIC), and the sample size-adjusted BIC (sBIC). These indices take model complexity (i.e., the number of parameters) into account and penalize highly parametrized models.

In order to test whether variables were associated with RGB, we performed multivariate Wald tests of multinomial logit regression weights (Equation 2). The first test served as a test of no association (NA), in which we simultaneously tested all weights attached to a covariate x_j against zero (i.e., $\omega_{11j} = \omega_{12j} = \omega_{13j} = \omega_{41j} = 0$). The second test was a test of constant associations (CA) and examined the equality of logistic regression weights (i.e., $\omega_{11j} = \omega_{12j} = \omega_{13j} = \omega_{41j}$). The CA test is interesting because it indicates whether the effects of covariates on RGB differ between regions (i.e., item positions) in the test. For example, if a covariate is significantly related to RGB (i.e., significant NA test), but the covariate's effects do not differ from each other (i.e., non-significant CA test), it implies that the covariate's effects on RGB constantly increase across item positions (i.e., the curves expected for two values of the covariates have similar shapes but different gradients). In contrast, a significant CA test indicates that the effects of a covariate do not constantly increase across positions, which means that the probability curves predicted at different values of the covariate differ in their shapes. For example, it might turn out that the effect of a covariate is limited to the first latent class ($C = 1$), whereas its effects on classes $C = 2$ and $C = 3$ are near to zero. Imagining this case, differences in RGB probabilities at different levels of the covariate would already arise early in the test session and would then remain constant across subsequent item positions. Alternatively, if the covariate's effects turn out to be stronger on class $C = 3$ and close to zero on classes $C = 2$ and $C = 1$, it means that the covariate's effects emerge only in the last section of the test. Hence, the CA test does not indicate a certain type of relationship. Instead, it indicates a non-constant pattern of relationships.

RESULTS

Unconditional LCA Models

In a first step we employed LCA models that did not include any covariates. The analyses served mainly descriptive purposes and were further used to evaluate the model's ability to depict the marginal RGB probabilities. Our proposed LCA model fitted the data better than a comparison model that assumed two classes (students with no or some RGB) in which the thresholds of all RGB indicators were unconstrained in the RGB class and estimated differently for text-only and text-picture items (unconstrained two-class model: #Parameters = 71, Log Likelihood = −2,313.5, AIC = 4,769.1, BIC = 5,052.3, sBIC = 4827.0; present model: #Parameters = 10, Log Likelihood = −1,963.1, AIC = 3,946.2, BIC = 3,986.0, sBIC = 3,954.3). This result indicates that our LCA model provided a good description of RGB. **Figure 1** presents the class-specific RGB probabilities by item position, uncovered by our LCA model, whereas the model fitted and observed RGB proportions are presented in the first panel of **Figure 2**. In line with previous results, the LCA model indicated that text-only items were more strongly affected by RGB ($\hat{\gamma} = -1.05$, $SE = 0.17$, $p < 0.001$). Furthermore, the LCA model categorized 56.6% of respondents as not engaging in RGB (observed data: 63.9%).

With respect to the onset of RGB, the LCA indicated that most students started to switch to this behavior in the later part of the test (23.8% in Class 3). The remaining classes had quite similar proportions, ranging between 5.0 and 8.0% (**Figure 1**). As can be seen in **Figure 2**, the five classes were sufficient for describing the marginal distribution of RGB for both text-only and text-picture items. Hence, the model appeared to be a solid starting point for assessing the predictors of RGB.

Next, we investigated changes in students' enjoyment ratings over the course of the test. We started with a linear growth curve model that was fitted to the data without considering the remaining variables. The model indicated that text-picture items were associated with higher enjoyment ratings throughout

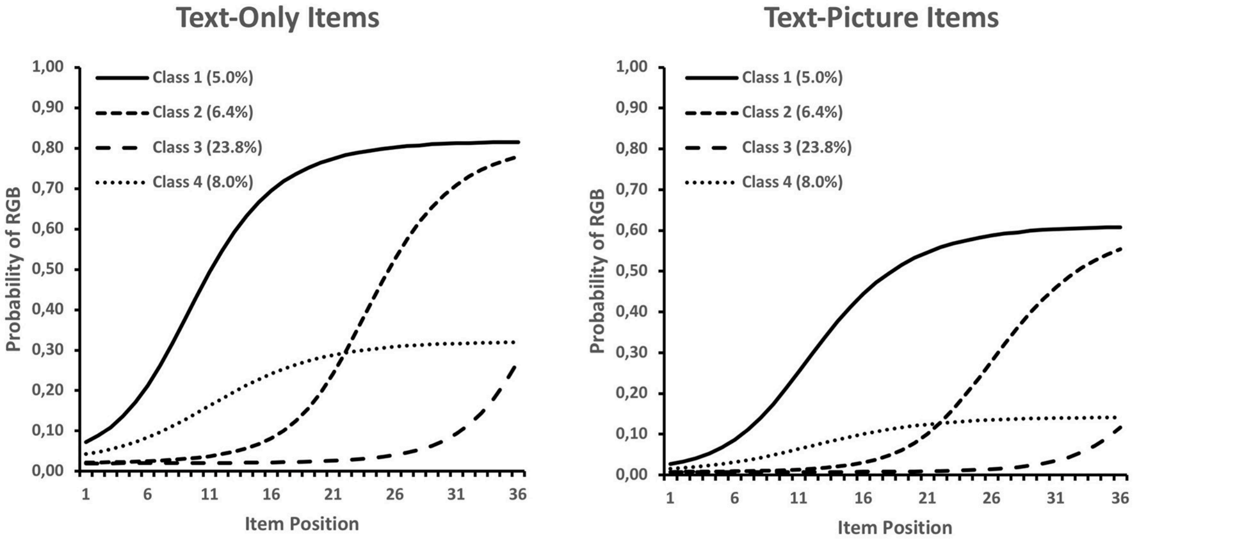


FIGURE 1 | RGB probability for latent Class 1 (early onset point), Class 2 (intermediate onset point), Class 3 (late onset point), and Class 4 (constantly low RGB) with results for text-only (left) and text-picture items (right).

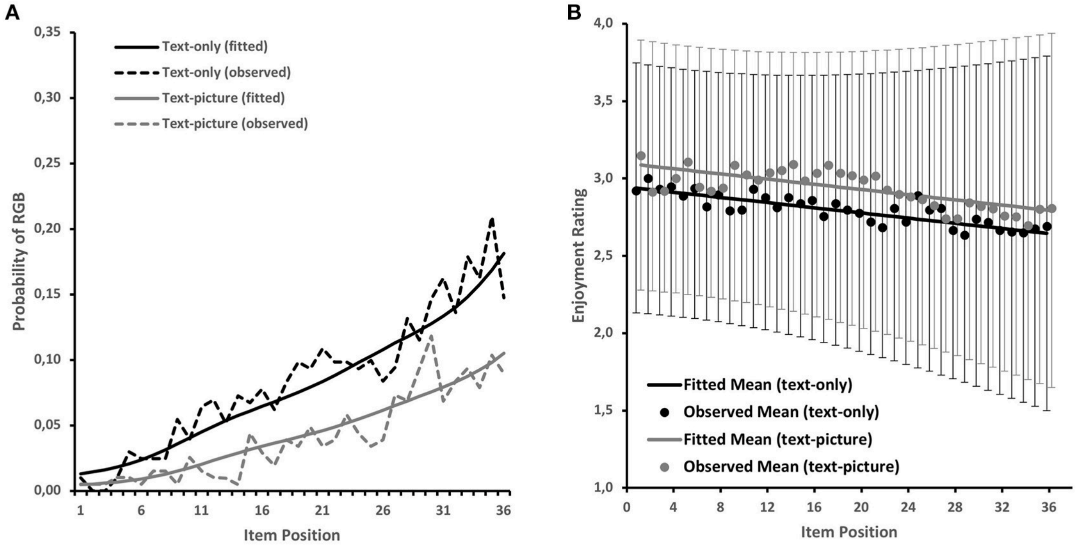


FIGURE 2 | (A) Observed and model-fitted RGB probabilities for text-only and text-picture items. **(B)** Observed (dots) and fitted (lines) average enjoyment ratings across item positions and distribution of fitted ratings (10th–90th percentiles) for text-only and text-picture items.

the test-taking session ($\hat{\delta} = 0.15$, $SE = 0.02$, $p < 0.001$), and that enjoyment ratings were, on average, high at the beginning of the test ($\hat{\mu}_{\eta0} = 2.94$, $SE = 0.04$, $p < 0.001$) but decreased on average across positions ($\hat{\mu}_{\eta_1} = -0.29$, $SE = 0.04$, $p < 0.001$). The results provide evidence for the existence of individual differences in initial enjoyment levels ($\hat{\sigma}^2_{\eta_0} = 0.39$, $SE = 0.03$, $p < 0.001$) and changes in enjoyment ($\hat{\sigma}^2_{\eta_1} = 0.52$, $SE = 0.05$, $p < 0.001$), with the two components being only weakly related ($\hat{\rho}_{\eta_0,\eta_1} = -0.12$, $SE = 0.06$, $p = 0.049$). Hence, the growth curve model indicated that, regardless of their initial enjoyment level, students exhibited relatively large individual differences in enjoyment declines. This aspect is visualized in **Figure 2B**, where the model-predicted average declines are depicted together with the observed means and the distribution of model-predicted scores (10th–90th percentiles of the distribution) that document increasing individual differences in enjoyment due to individual differences in the trajectories.

Conditional LCA Models

To study the correlates of RGB, we started by employing conditional LCA models in which we used each predictor in isolation without considering the remaining covariates. The exceptions were the two latent variables of the growth curve model applied to the enjoyment variables that were investigated simultaneously. **Table 1** presents multinomial regression weights

TABLE 1 | Multinomial logistic regression weights determined separately for each covariate, and corresponding Wald-χ^2 tests of no association (*NA*) and of constant associations (*CA*).

	Gender	School type	General cognitive abilities	Working memory capacity	Initial task enjoyment	Change in task enjoyment
	Est. (SE)	**Est. (SE)**	**Est. (SE)**	**Est. (SE)**	**Est. (SE)**	**Est. (SE)**
C = 1	−0.70 (0.51)	−4.25 (1.06)**	−1.02 (0.20)**	−2.11 (0.40)**	−0.84 (0.52)	−0.91 (0.40)*
C = 2	−0.75 (0.50)	−10.57 (1.30)**	−0.80 (0.22)**	−0.45 (0.33)	−0.88 (0.36)*	−1.27 (0.40)**
C = 3	−0.32 (0.32)	−1.23 (0.33)**	−0.13 (0.21)	−0.30 (0.20)	−0.20 (0.30)	−0.37 (0.34)
C = 4	−0.81 (0.50)	−3.03 (0.91)**	−1.11 (0.24)**	−1.45 (0.39)**	−0.72 (0.35)*	0.13 (0.34)
	χ^2 (df)	**χ^2 (df)**	**χ^2 (df)**	**χ^2 (df)**	**χ^2 (df)**	**χ^2 (df)**
NA	7.05 (4)	113.20 (4)**	44.37 (4)**	31.13 (4)**	11.42 (4)*	16.07 (4)**
CA	1.32 (3)	59.30 (3)**	20.00 (3)**	18.18 (3)**	3.61(3)	10.94(3)*

*Gender: 0 = male, 1 = female; school type: 0 = academic track (Gymnasium), 1 = non-academic track (i.e., regional school); Measures of general cognitive abilities (KFT) and working-memory were standardized prior to the analysis. *$p \leq 0.05$; **$p \leq 0.01$.*

determined for each variable and the corresponding tests for no association (row NA) and constant associations (row CA) with RGB.

As can be seen in **Table 1**, almost all variables were significantly related to RGB. The exception was gender. The pattern of gender differences was in line with previous results but did not reach the significance threshold ($p = 0.113$). Judged on the value of the Wald-χ^2 statistic, school type was most strongly related to RGB, whereas the initial level of and change in enjoyment had the weakest relationships to RGB. Furthermore, the four multinomial logistic regression weights belonging to each variable appeared to differ from each other. For example, the regression weights associated with school type indicated that the chances of academic-track students belonging to classes $C = 1$, 2, or 4 vs. class $C = 5$ were much smaller than the corresponding chances of non-academic-track students. In contrast, school-type differences in the relative chance of belonging to class $C = 3$ (i.e., the late RGB onset class) were less pronounced (i.e., the regression weight was closer to zero).

As can be seen in the CA row in **Table 1**, school type was differentially related to the onset point of RGB, whereas gender and initial enjoyment were not. Academic-track students were least likely to have an early RGB onset (i.e., membership in classes $C = 1$, 2, or 4). Similar relationships were found with the continuous covariates, general cognitive abilities, working-memory capacity, and change in enjoyment, so that students with higher scores on these variables were least likely to have an early RGB onset.

In order to get an impression of the pattern of relationships, the model-predicted probabilities of RGB at selected values of the covariates are plotted in **Figure 3**. As suggested by the non-significant overall effect (NA test, **Table 1**) and the non-significant CA test, gender differences were rather small, but showed a relatively constant (albeit non-significant, $p = 0.113$) increase across positions. In contrast, differences between school types were clearly larger and showed a strong increase across item positions, whereby the increase was largest in the first two thirds of the test. A similar picture was revealed for the continuous measures of general cognitive abilities and working-memory capacity. In the case of these variables, it appeared that above average scores did not have a meaningful effect on RGB. Rather, students who scored well below average on these tests had a higher probability of engaging in RGB.

The relationship of RGB with the repeatedly measured enjoyment variable is shown in **Figure 4**. In order to account for the initial level and the change component in the enjoyment ratings, the figure contains three line plots for low (10th percentile), average, and high levels (90th percentile) of initial enjoyment, which each contain RGB probability curves for low (10th percentile), average, and high levels (90th percentile) of change in enjoyment. As shown in **Figure 4**, lower initial levels of enjoyment were associated with constantly increasing levels of RGB across positions (non-significant CA test). As further shown in **Figure 4**, the RGB probability curves differed at each level of initial enjoyment, depending on the change in enjoyment, so that steeper decreases in enjoyment were associated with steeper increases in RGB (see also NA row in **Table 1**).

All results presented up to this point pertain to the models in which each covariate was investigated in isolation. However, the majority of student characteristics employed were correlated among each other, as can be taken from **Table 2**. Even though the correlations were not so high that they could cause collinearity problems, the question about each variable's unique contribution to the prediction of RGB emerged. We approached this question by using all covariates simultaneously as predictors of latent class membership. The results are presented in **Table 3**.

The (non-significant) relationship of gender with RGB was not affected by the inclusion of the other covariates (see **Table 1**). A similar result was found for school type; RGB was still significantly related to this variable and also strongly related to an early RGB onset. The relationship of general cognitive abilities with RGB was clearly reduced after all covariates were included in the model, although the relationship with RGB and RGB onset remained significant. In contrast, the relationship of working memory with RGB was similar to that of the previous model (see **Table 1**), which means that it continued to be significantly related to RGB and its onset. Initial enjoyment also remained significantly related to RGB, but the regression weights for the different latent classes did not differ significantly (CA test;

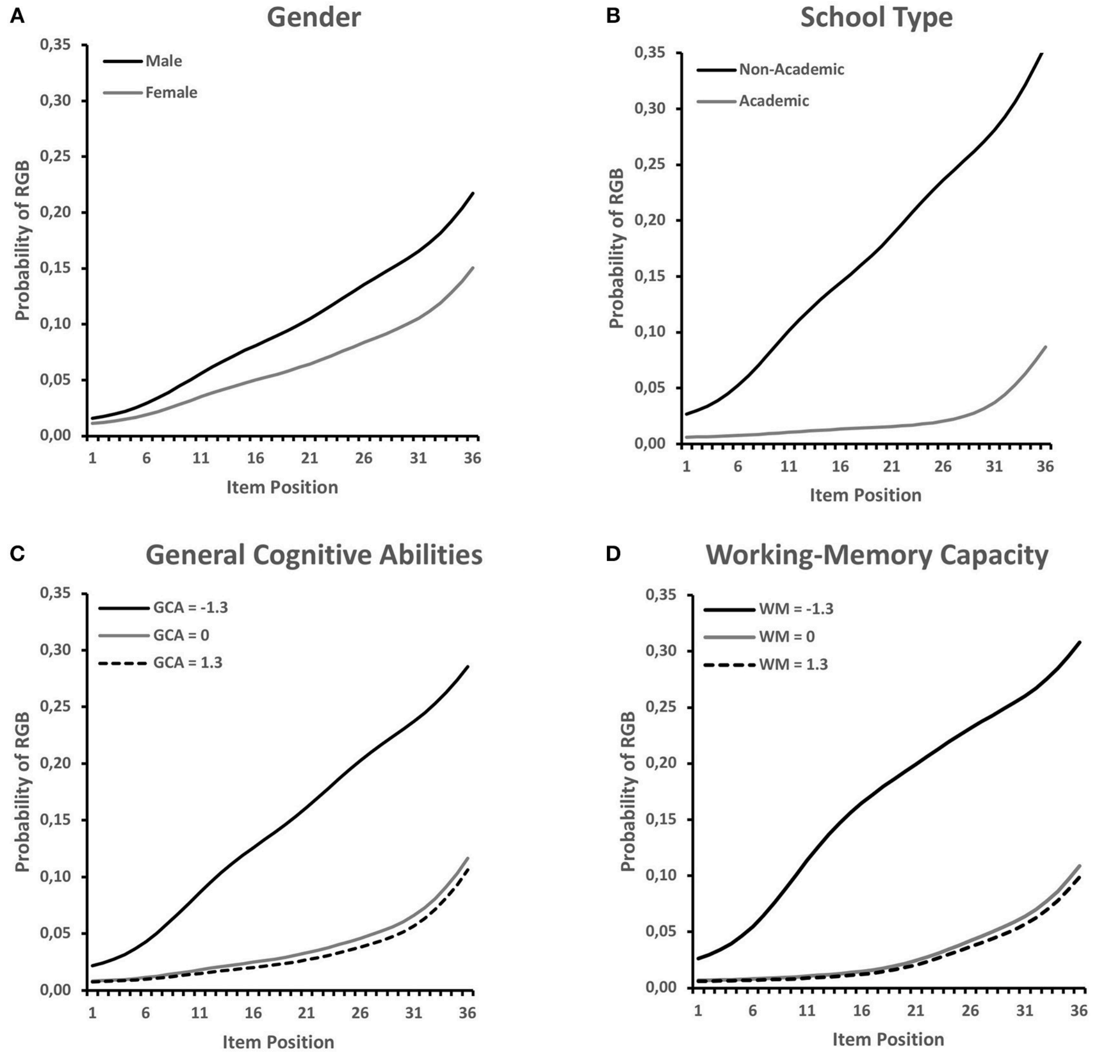


FIGURE 3 | Estimated RGB probabilities by item position expected for different levels of the covariates **(A)** gender, **(B)** school type, **(C)** general cognitive abilities, and **(D)** working-memory capacity. Values ±1.3 standard deviations around the mean were chosen for general cognitive abilities and working-memory capacity because these roughly indicate the 10th and 90th percentiles of their distribution.

see **Table 3**). Finally, changes in enjoyment continued to be significantly related to RGB, but the CA test was no longer significant on the $p < 0.05$ level ($p = 0.054$). This weakens the evidence of a strong relation between students' enjoyment decline and early RGB onset.

DISCUSSION

The present study examined the correlates of RGB onset and its temporal dynamic over the course of testing as a between-student factor with regard to motivational and cognitive student characteristics, using a latent class approach as a base for our analyses. Specifically, we investigated the extent to which different patterns of (early) RGB onset were related to cognitive and motivational covariates in order to gain deeper insights into the processes that may underlie disengaged test-taking behavior in low-stakes assessment. In the following sections, we discuss the key results of the study with regard to our hypotheses, the theoretical assumptions, and earlier research. Finally, we reflect on the study's limitations, consider future research suggestions, and close the article with an overall conclusion and a consideration of the practical significance of our findings.

Student Characteristics

Testing our hypothesis regarding the relation of RGB or RGB onset and students' gender (H1a), we did not find a significant relation, contrary to our expectation. However, this is not entirely surprising, as the findings in the literature are also inconsistent. Several studies indicate that male students have lower levels of test-taking motivation and also tend to show disengaged

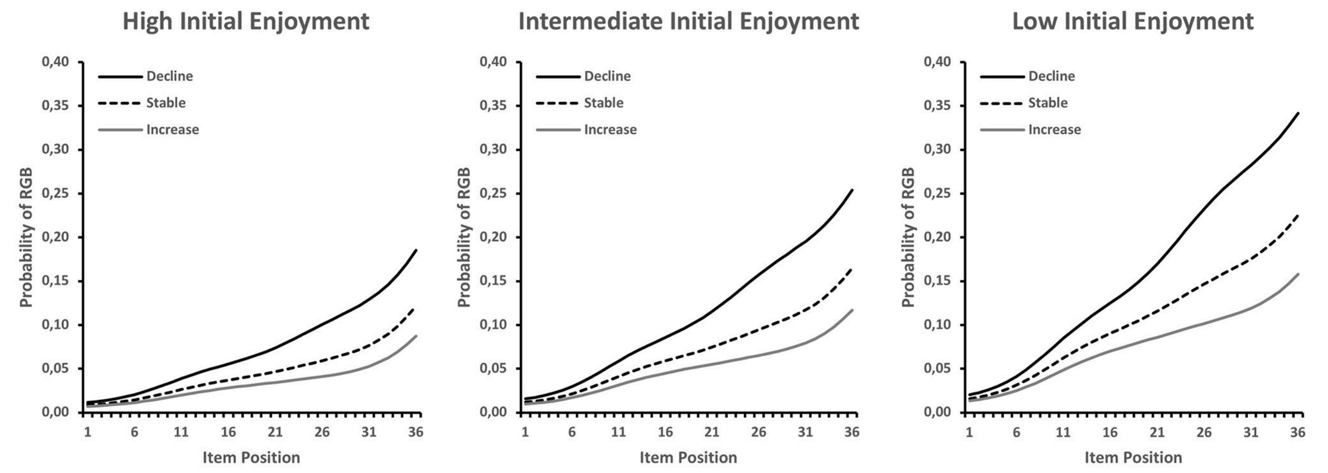


FIGURE 4 | RGB probabilities for text-only items by item position, expected for different combinations of initial enjoyment and change in enjoyment over the course of the test (i.e., item positions).

TABLE 2 | Predictor correlations.

	1	2	3	4	5	6
1. Gender	1					
2. School type	−0.039	1				
3. General cognitive abilities	−0.002	0.363**	1			
4. Working memory	0.025	0.298**	0.281**	1		
5. Initial task enjoyment	0.024	0.066	0.060	−0.028	1	
6. Change in task enjoyment	0.086	0.076	0.099	0.039	−0.123*	1

*$p \leq 0.05$; **$p \leq 0.01$.

behavior, such as RGB, more often (for a review see e.g., DeMars et al., 2013). Still, not all studies find a significant relation between gender and RGB (e.g., Wise et al., 2009). In the present study, as can be seen in **Figure 3A**, the descriptive pattern was in line with the expectation that male students would engage in RGB earlier than female students, but the coefficient did not reach significance. This result seems to be primarily related to a power issue, as the present sample may not have been large enough to significantly show the effect. Generally, the relationship between RGB and gender appeared to be of lower practical importance considering the marginal effect sizes in vast representative samples, such as in the study by Goldhammer et al. (2016). However, gender differences in RGB may be more pronounced in younger students, which seemed to be reflected at a descriptive level in our data. The moderating role of students' age would, thus, be an interesting factor for future research.

Confirming our hypothesis regarding students' school-type attendance (H1b), we found a remarkably higher risk of an earlier RGB onset and a stronger increase of RGB probabilities in students from non-academic-track schools (see **Figure 3B**). This effect remained significant when all predictors were included in one model; moreover, school type was the strongest predictor of early RGB onset. In the German school system, which assigns students to different secondary school tracks based on their performance in elementary school, school type is strongly related to students' academic abilities (e.g., Prenzel et al., 2013). In addition, school type has been shown to be connected to differences in students' motivation to work in an effortful way in low-stakes assessments (e.g., Baumert and Demmrich, 2001; Nagy et al., 2018b). Thus, both factors, academic ability and motivation, are probably reflected in the substantial RGB differences between school tracks. Earlier studies have shown similar relations of RGB (e.g., Lee and Jia, 2014; Goldhammer et al., 2016; Wise and Gao, 2017) or item position effects (e.g., Nagy et al., 2016, 2018a) with students' academic ability level (e.g., SAT scores; Wise et al., 2009) or school-type attendance (Nagy et al., 2016). Nevertheless, some studies did not find ability-related differences in students' response effort (e.g., Wise and DeMars, 2005; Wise and Kong, 2005; Wise et al., 2006a). These mixed results might be attributed to the different sample characteristics, test situations, and criteria used to judge students' academic ability (e.g., scores from the investigated test vs. external criteria, such as SAT scores). In this study, we used a criterion that is independent of students' test achievement and known to be a solid indicator of academic abilities. However, while the investigated data set included students from academic- and non-academic-track schools, it did not reflect the full width of German non-academic-track schools (i.e., no lower secondary schools). Our findings might therefore not fully represent school-type differences, as students from lower non-academic schools might further contribute to the unfavorable picture of school-type differences in RGB.

In line with our hypotheses regarding students' general cognitive abilities (H1c) and working-memory capacity (H1d), we found substantial evidence that both factors are significantly related to RGB and predict an earlier RGB onset and a stronger increase in RGB. However, this only applied to students with relatively low cognitive capacities (see **Figures 3C,D**). This indicates that a lack of cognitive resources raises students' risk of engaging in RGB early on and of showing a stronger RGB increase. Building on expectancy-value models (e.g., Eccles and Wigfield, 2002) and the DCM assumptions (Wise, 2017), this

TABLE 3 | Multinomial logistic regression weights determined jointly for all covariates, and corresponding Wald-χ^2 tests of no association (*NA*) and of constant associations (*CA*).

	Gender	School type	General cognitive abilities	Working-memory capacity	Initial task enjoyment	Change in task enjoyment
	Est. (SE)	**Est. (SE)**	**Est. (SE)**	**Est. (SE)**	**Est. (SE)**	**Est. (SE)**
C = 1	−0.84 (0.79)	−4.77 (0.97)**	−0.60 (0.36)	−2.54 (0.69)**	−1.39 (0.67)*	−1.22 (0.68)
C = 2	−0.86 (0.75)	−3.30 (0.37)**	−0.38 (0.21)	−0.26 (0.42)	−1.18 (0.49)*	−1.12 (0.50)*
C = 3	−0.59 (0.33)	−1.37 (0.43)**	0.13 (0.20)	−0.22 (0.20)	−0.34 (0.32)	−0.29 (0.32)
C = 4	−1.07 (0.69)	−3.00 (0.77)**	−0.80 (0.30)**	−1.24 (0.49)**	−1.10 (0.56)	0.02 (0.47)
	χ^2 (df)	**χ^2 (df)**	**χ^2 (df)**	**χ^2 (df)**	**χ^2 (df)**	**χ^2 (df)**
NA	6.14 (4)	99.94 (4)**	9.59 (4)*	17.69 (4)**	11.52 (4)*	10.12 (4)*
CA	0.68 (3)	33.85 (3)**	9.15 (3)*	12.07 (3)**	5.27 (3)	7.63 (3)

Gender: 0 = male, 1 = female; school type: 0 = academic track (Gymnasium), 1 = non-academic track (i.e., regional school); Measures of general cognitive abilities (KFT) and working-memory were standardized prior to the analysis. *p ≤ 0.05; **p ≤ 0.01.

is not really surprising. However, so far, we are not aware of any empirical studies that have investigated standardized cognitive ability tests as predictors of RGB development so far. While both of our measures were clearly related to RGB as isolated predictors, it is especially interesting that working-memory capacity seemed to be more predictive of both RGB and RGB onset than general cognitive abilities. This became evident when we integrated all indicators into one full competitive model, where the general cognitive ability covariate lost a substantial part of its explanatory power but the working-memory factor remained basically unaffected. This might be explained as follows: Whereas general cognitive abilities are assumed to be more or less stable across situations and time (i.e., fluid intelligence as a trait), working-memory capacity is known to be subject to stronger situational fluctuations (see e.g., Hofmann et al., 2011) and can also be subject to mental fatigue effects that undermine attentional control (Schmeichel, 2007). However, executive attention is a key factor in self-controlled behavior, which is also needed in any test situation in order for students to focus on the posed problems and to solve them with effort. This demand tends to become aversive over the course of testing time (Inzlicht et al., 2014). This relation could help to explain why working-memory capacity seems to be the more important cognitive resource required for engaged test-taking behavior over the course of a test session.

Task Enjoyment Over the Course of Testing

RGB is typically interpreted as an indicator of student motivation. In our study, we examined the extent to which RGB was related to students' perceived motivation level as an open question. By modeling the intercept of students' multiple enjoyment ratings across the test session as a latent covariate in the LCA, we tested Hypothesis H2a. Although there was evidence for a relation between students' initial enjoyment (i.e., rating of the first item) and RGB, we did not find a significant relation to RGB onset (**Figure 4**). This was true for both the isolated analysis of initial enjoyment as a single predictor and the full model with all predictors. The observed and model-fitted data of students' enjoyment ratings (**Figure 2B**) showed a decrease over the course of the test session, as expected, though the mean level of students' enjoyment remained relatively high. The figure also shows that there was a lot of inter-individual variance; we investigated this variance by integrating students' estimated slopes as a latent covariate into our LCA to test Hypothesis H2b. This provided tentative evidence that a negative enjoyment trajectory over time predicted both RGB and RGB onset in the isolated model. However, the relation with RGB onset did not remain significant when competitive covariates were added to the model, which weakens the evidence for Hypothesis H2b to some extent.

Overall, students' enjoyment ratings were not strongly related to their RGB tendency when compared to the cognitive covariates. This relatively weak relation could be due to the young age of the students in the current sample, who might not yet be able to correctly reflect on their current enjoyment; but, it could also indicate that test-takers simply have problems with an accurate evaluation of their motivational state. However, this question cannot be answered based on the present findings. Penk and Richter (2017) recently applied a comparable approach of modeling ninth-graders' test-taking motivation across a test session to investigate item position effects. They found that initial test-taking motivation was a better predictor of the item position effect than changes in motivation. This pattern is the opposite of our results and is somewhat surprising; it indicates that there are interesting questions to be answered in future research on test-taking motivation.

Limitations and Future Directions

Some limitations need to be taken into account when interpreting the present findings. First, the current sample cannot be considered representative, which constrains the generalizability. The effects of school type might be biased because we did not include all German school tracks and we tested only students in the fifth and sixth grades. Compared to typical large-scale assessments, the current sample was rather small but seemed to be sufficient, except for determining the relation between RGB and gender, which may have been underpowered. As an unusual advantage, however, the data included important measures, such as the repeated enjoyment rating and the indicators of students' general cognitive abilities and working-memory capacity, which were at the core of the present analyses. The test circumstances were highly comparable to typical computer-based low-stakes testing programs. Nevertheless, future studies should challenge our research and try to replicate the current findings in larger data sets. Especially a transfer of our latent class approach to other samples would be desirable to evaluate the extent to which the presumptions and findings of our study (e.g., the proportion of student assignments to the five individual LCA classes) are robust. As such, the proposed analysis could be a fruitful base for future research on the determinants of RGB onset and its dynamics across testing time.

Second, the reliability of our working-memory test (i.e., reversed digit span) was, unfortunately, not very high ($\alpha = 0.65$). However, a tradeoff has to be made with view to the challenge of measuring working-memory indicators in group sessions, as individual test sessions can better ensure that the test is administered in the best way possible. It would therefore be advantageous to reexamine the current issue by assessing other or additional working-memory capacity indicators that have a higher test reliability.

A third potential limitation pertains to the fact that both the science test and the cognitive tests (KFT N2 and reverse digit span) were administered in the same test session. The results might therefore share common variance due to a general tendency of students to work seriously on test items in a low-stakes situation (i.e., in terms of a latent trait) and also due to their current overall compliance with the test-taking situation (i.e., in terms of a current state during the specific test administration). However, the cognitive tests were presented before the science test. The risk that students' behavior was already effortless at the beginning of the test session is rather low. This assumption is supported by the observation that only a small number of RGB trials occurred in the first items of the science test, indicating that most students were still prepared to make an effort to work on the test items at the beginning of the test. Nevertheless, test scores from standardized cognitive tests that were assessed in different sessions from another day would have been preferable.

Conclusion and Implications for Educational Practice

Drawing on a theory-driven latent class model, standardized measures of students' cognitive abilities, and repeated ratings of their current item-solving enjoyment, this study was able to extend previous work and widen the understanding of RGB. The main strength of our investigation is that our LCA approach made it possible to study the dynamics of RGB in connection with several indicators of cognitive and motivational resources at a student level. In brief, we found evidence that students' item-solving enjoyment, academic ability, and cognitive capacities are (closely) related to the RGB onset point and the dynamics of RGB across a low-stakes test session. Students from non-academic-track schools, students with low general cognitive abilities and low working-memory capacity, as well as students with a stronger decline in their task enjoyment over the course of the test were substantially more likely to engage in RGB earlier in the test and to progress with that behavior. All of these findings are in line with the theoretical assumptions from expectancy-value models (e.g., Eccles and Wigfield, 2002) as well as those of the DCM by Wise and Smith (2011). However, future research should also focus on non-cognitive factors, such as coping strategies, text anxiety or the well-being of students and on the relations of these factors to test-taking behavior. In addition, characteristics of students' home environment, such as the socio-economic status of their parents and school culture, including the school climate, the ethnic composition and the value teachers, parents and peers attribute to learning and testing efforts, should be taken into account in order to better understand RGB from a broader perspective.

Alongside the new support they provide for the theoretical models concerning the psychological determinants of RGB, our results also have practical implications. The substantial relation of RGB to students' academic and cognitive abilities suggests that students' test engagement seems to be a seriously, confounding factor (in terms of true competences) for a valid interpretation of school-type comparisons of low-stakes test performances (see also Wise et al., 2009; Nagy et al., 2018b). This is a problem because such comparisons are often an important goal of large-scale testing programs. Furthermore, all motivation-filtering procedures rely on the theoretical assumption that student motivation is unrelated to true proficiency. However, if this criterion is not fulfilled, the filtering procedure induces bias. In particular, filtering students with low proficiency out of the data would provide an overly positive picture of the performance in the investigated sample, leading to an overestimation of true proficiency. In addition to the attempt of using statistical correction procedures, this problem should also be discussed at the level of test characteristics. For example, applying shorter tests, using items with a more appealing design (see e.g., Lindner et al., 2017b, Wise et al., 2009), and possibly having longer breaks between different test blocks may foster students' test-taking motivation and could allow them to refresh exhausted cognitive resources before continuing to focus their attention on further tasks (see also Lindner et al., 2018). In the light of the current results, such considerations seem to be particularly relevant for students from non-academic-track schools and for students with low working-memory capacity. However, the extent to which an improvement in assessment conditions would actually contribute to solving the problems that are connected to low test effort is a question for future research.

ETHICS STATEMENT

This study was carried out in accordance with the Declaration of Helsinki and the ethical guidelines for experimental research with human participants as proposed by the German Psychological Society (DGPs). Prior to the test session, we obtained written informed consent from all students and their legal guardians.

AUTHOR CONTRIBUTIONS

All authors have made a substantial intellectual contribution to the article and approved it for publication. ML developed the study design, performed the data collection, prepared the data for analyses and wrote the manuscript. GN developed the conception of the analyses, performed the analyses and co-wrote the statistical analyses paragraph. GN and OL reviewed and edited the article. All authors contributed to the theoretical conception, the interpretation of the results and manuscript revisions.

ACKNOWLEDGMENTS

This work was funded by the Leibniz Institute for Science and Mathematics Education (IPN) in Kiel.

REFERENCES

Ackerman, P. L., and Kanfer, R. (2009). Test length and cognitive fatigue: an empirical examination of effects on performance and test-taker reactions. *J. Exp. Psychol. Appl.* 15, 163–181. doi: 10.1037/a0015719

Ackerman, P. L., Kanfer, R., Shapiro, S. W., Newton, S., and Beier, M. E. (2010). Cognitive fatigue during testing: an examination of trait, time-on-task, and strategy influences. *Hum. Perform.* 23, 381–402. doi: 10.1080/08959285.2010.517720

Barry, C. L., and Finney, S. J. (2016). Modeling change in effort across a low-stakes testing session: a latent growth curve modeling approach. *Appl. Meas. Educ.* 29, 46–64. doi: 10.1080/08957347.2015.1102914

Baumert, J., and Demmrich, A. (2001). Test motivation in the assessment of student skills: the effects of incentives on motivation and performance. *Eur. J. Psychol. Educ.* 16, 441–462. doi: 10.1007/BF03173192

Cronbach, L. J. (1960). *Essentials of Psychological Testing, 2nd Edn*. New York, NY: Harper & Row.

Debeer, D., Buchholz, J., Hartig, J., and Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *J. Educ. Behav. Stat.* 39, 502–523. doi: 10.3102/1076998614558485

DeMars, C. E., Bashkov, B. M., and Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Res. Pract. Assess.* 8, 69–82. Available online at: http://www.rpajournal.com/dev/wp-content/uploads/2013/11/A4.pdf

Eccles, J. S., Adler, T. E., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., et al. (1983). "Expectancies, values, and academic behaviors," in *Achievement and Achievement Motivation*, ed J. T. Spence (San Francisco, CA: W.H. Freeman), 75–146.

Eccles, J. S., and Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annu. Rev. Psychol.* 53, 109–132. doi: 10.1146/annurev.psych.53.100901.135153

Engle, R. W. (2002). Working memory capacity as executive attention. *Psychol. Sci.* 11, 19–23. doi: 10.1111/1467-8721.00160

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Res. Rep. Ser.* 2015, 1–17. doi: 10.1002/ets2.12067

Goldhammer, F., Martens, T., Christoph, G., and Lüdtke, O. (2016). *Test-Taking Engagement in PIAAC, OECD Education Working Papers, No. 133.* Paris: OECD Publishing. doi: 10.1787/5jlzfl6fhxs2-en

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., and Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: insights from a computer-based large-scale assessment. *J. Educ. Psychol.* 106, 608–626. doi: 10.1037/a0034716

Greiff, S., Wüstenberg, S., and Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Comput. Educ.* 91, 92–105. doi: 10.1016/j.compedu.2015.10.018

Haladyna, T. M., and Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educ. Meas. Issues Pract.* 23, 17–27. doi: 10.1111/j.1745-3992.2004.tb00149.x

Hartenstein, S. (2012). *flexSURVEY – Flexible PHP-Driven Online Surveys (Version 2.0)* [Computer software]. Retrieved from http://www.flexsurvey.de/Download

Hartig, J., and Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychol. Test Assess. Model.* 54, 418–431. Available online at: http://journaldatabase.info/articles/multilevel_item_response_model_for.html

Heller, K. A., and Perleth, C. (2000). *KFT 4–12+ R: Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision. [Cognitive Abilities Test for Students from Grade 4 to 12+ (CogAT; Thorndike, L. & Hagen, E., 1954–1986) German adapted version/author]*. Göttingen: Beltz.

Hofmann, W., Schmeichel, B. J., Friese, M., and Baddeley, A. D. (2011). "Working memory and self-regulation," in *Handbook of Self-Regulation*, eds K. D. Vohs and R. F. Baumeister (New York, NY; London: The Guilford Press), 204–225.

International Association for the Evaluation of Educational Achievement [IEA] (2013). *TIMSS 2011 Assessment Released Science Items.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College/International Association for the Evaluation of Educational Achievement (IEA), IEA Secretariat. Retrieved from http://nces.ed.gov/timss/pdf/TIMSS2011_G4_Science.pdf

Inzlicht, M., Schmeichel, B. J., and Macrae, C. N. (2014). Why self-control seems (but may not be) limited. *Trends Cogn. Sci.* 18, 127–133. doi: 10.1016/j.tics.2013.12.009

Jin, K. Y., and Wang, W. C. (2014). Item response theory models for performance decline during testing. *J. Educ. Meas.* 51, 178–200. doi: 10.1111/jedm.12041

Kong, X. J., Wise, S. L., and Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educ. Psychol. Meas.* 67, 606–619. doi: 10.1177/0013164406294779

Lee, Y. H., and Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assess. Educ.* 2, 1–24. doi: 10.1186/s40536-014-0008-1

Lindner, C., Nagy, G., Ramos Arhuis, W. A., and Retelsdorf, J. (2017a). A new perspective on the interplay between self-control and cognitive performance: modeling progressive depletion patterns. *PLoS ONE* 12:e0180149. doi: 10.1371/journal.pone.0180149

Lindner, C., Nagy, G., and Retelsdorf, J. (2018). The need for self-control in achievement tests: changes in students' state self-control capacity and effort investment. *Soc. Psychol. Educ.* 21, 1113–1131. doi: 10.1007/s11218-018-9455-9

Lindner, M. A., Ihme, J. M., Saß, S., and Köller, O. (2016). How representational pictures enhance students' performance and test-taking pleasure in low-stakes assessment. *Eur. J. Psychol. Assess.* 34, 376–385. doi: 10.1027/1015-5759/a000351

Lindner, M. A., Lüdtke, O., Grund, S., and Köller, O. (2017b). The merits of representational pictures in educational assessment: evidence for cognitive and motivational effects in a time-on-task analysis. *Contemp. Educ. Psychol.* 51, 482–492. doi: 10.1016/j.cedpsych.2017.09.009

List, M. K., Robitzsch, A., Lüdtke, O., Köller, O., and Nagy, G. (2017). Performance decline in low-stakes educational assessments: different mixture modeling approaches. *Large-scale Assess. Educ.* 5:15. doi: 10.1186/s40536-017-0049-3

Lu, H., Tian, Y., and Wang, C. (2018). The influence of ability level and big five personality traits on examinees' test-taking behaviour in computerised adaptive testing. *Int. J. Soc. Media Interact. Learn. Environ.* 6, 70–84. doi: 10.1504/IJSMILE.2018.092380

Messick, S. (1989). Meaning and values in test validation: the science and ethics of assessment. *Educ. Res.* 18, 5–11. doi: 10.3102/0013189X018002005

Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., and Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks*. Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

Muthén, L. K., and Muthén, B. O. (2017). *Mplus user's guide, 8th Edn*. Los Angeles, CA: Muthén & Muthén.

Nagy, G., Lüdtke, O., and Köller, O. (2016). Modeling test context effects in longitudinal achievement data: examining position effects in the longitudinal German PISA 2012 assessment. *Psychol. Test Assess. Model.* 58, 641–670.

Nagy, G., Nagengast, B., Becker, M., Rose, N., and Frey, A. (2018a). Item position effects in a reading comprehension test: an IRT study of individual differences and individual correlates. *Psychol. Test Assess. Model.* 60, 165–187. Available online at: https://www.psychologie-aktuell.com/fileadmin/download/ptam/2-2018_20180627/03_PTAM-2-2018_Nagy_v2.pdf

Nagy, G., Nagengast, B., Frey, A., Becker, M., and Rose, N. (2018b). A multilevel study of position effects in PISA achievement tests: student- and school-level predictors in the German tracked school system. *Assess. Educ. Princ. Policy Pract. Adv.* 60: 165–187. doi: 10.1080/0969594X.2018.1449100

Penk, C., and Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educ. Assess. Eval. Account.* 29, 55–79. doi: 10.1007/s11092-016-9249-6

Petermann, F., and Petermann, U. (Eds). (2010). *HAWIK-IV: Hamburg-Wechsler-Intelligenztest für Kinder-IV; Manual; Übersetzung und Adaption der WISC-IV von David Wechsler.* Bern: Huber.

Prenzel, M., Sälzer, C., Klieme, E., and Köller, O. (Eds). (2013). *PISA 2012. Fortschritte und Herausforderungen in Deutschland [PISA 2012. Advances and challenges in Germany]*. Münster: Waxmann.

Rios, J. A., Guo, H., Mao, L., and Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: to filter unmotivated examinees or not? *Int. J. Test.* 17, 74–104. doi: 10.1080/15305058.2016.1231193

Schmeichel, B. J. (2007). Attention control, memory updating, and emotion regulation temporarily reduce the capacity for executive control. *J. Exp. Psychol. Gen.* 136, 241–255. doi: 10.1037/0096-3445.136.2.241

Setzer, J. C., Wise, S. L., van den Heuvel, J., and Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Appl. Meas. Educ.* 26, 34–49. doi: 10.1080/08957347.2013.739453

Swerdzewski, P. J., Harmes, J. C., and Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Appl. Meas. Educ.* 24, 162–188. doi: 10.1080/08957347.2011.555217

Thelk, A. D., Sundre, D. L., Horst, S. J., and Finney, S. J. (2009). Motivation matters: using the student opinion scale to make valid inferences about student performance. *J. Gen. Educ.* 58, 129–151. doi: 10.1353/jge.0.0047

Weirich, S., Hecht, M., Penk, C., Roppelt, A., and Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Appl. Psychol. Meas.* 41, 115–129. doi: 10.1177/0146621616676791

Wigfield, A., and Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemp. Educ. Psychol.* 25, 68–81. doi: 10.1006/ceps.1999.1015

Wise, S. L. (2006). An investigation of the differential effort received by items on a low stakes computer-based test. *Appl. Meas. Educ.* 19, 95–114. doi: 10.1207/s15324818ame1902_2

Wise, S. L. (2017). Rapid-guessing behavior: its identification, interpretation, and implications. *Educ. Meas. Issues Pract.* 36, 52–61. doi: 10.1111/emip.12165

Wise, S. L., Bhola, D., and Yang, S. (2006a). Taking the time to improve the validity of low-stakes tests: the effort-monitoring CBT. *Educ. Meas. Issues Pract.* 25, 21–30. doi: 10.1111/j.1745-3992.2006.00054.x

Wise, S. L., and DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. *Educ. Assess.* 10, 1–17. doi: 10.1207/s15326977ea1001_1

Wise, S. L., and DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educ. Assess.* 15, 27–41. doi: 10.1080/10627191003673216

Wise, S. L., and Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Appl. Meas. Educ.* 30, 343–354. doi: 10.1080/08957347.2017.1353992

Wise, S. L., and Kong, X. J. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Appl. Meas. Educ.* 18, 163–183. doi: 10.1207/s15324818ame1802_2

Wise, S. L., and Ma, L. (2012). Setting response time thresholds for a CAT item pool: The normative threshold method. *Paper Presented at the 2012 Annual Meeting of the National Council on Measurement in Education* (Vancouver, BC).

Wise, S. L., Ma, L., Kingsbury, G. G., and Hauser, C. (2010). An investigation of the relationship between time of testing and test-taking effort. *Paper Presented at the 2010 Annual Meeting of the National Council on Measurement in Education* (Denver, CO).

Wise, S. L., Pastor, D. A., and Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: implications for test development and measurement practice. *Appl. Meas. Educ.* 22, 185–205. doi: 10.1080/08957340902754650

Wise, S. L., and Smith, L. F. (2011). "A model of examinee test-taking effort," in *High-Stakes Testing in Education: Science and Practice in K-12 Settings*, eds J. A. Bovaird, K. F. Geisinger, and C. W. Buckendal (Washington, DC: American Psychological Association), 139–153. doi: 10.1037/12330-009

Wise, V. L., Wise, S. L., and Bhola, D. S. (2006b). The generalizability of motivation filtering in improving test score validity. *Educ. Assess.* 11, 65–83. doi: 10.1207/s15326977ea1101_3

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 20 September 2019
doi: 10.3389/fpsyg.2019.01808

The u-can-act Platform: A Tool to Study Intra-individual Processes of Early School Leaving and Its Prevention Using Multiple Informants

Frank J. Blaauw[1,2*†], *Mandy A. E. van der Gaag*[1†], *Nick R. Snell*[1], *Ando C. Emerencia*[1], *E. Saskia Kunnen*[1] *and Peter de Jonge*[1]

[1] Department of Developmental Psychology, Faculty Behavioral and Social Sciences, University of Groningen, Groningen, Netherlands, [2] Distributed Systems Group, Bernoulli Institute for Mathematics, Computer Science, and Artificial Intelligence, University of Groningen, Groningen, Netherlands

OPEN ACCESS

Edited by:
Frank Goldhammer,
German Institute for International
Educational Research (LG), Germany

Reviewed by:
Kathleen Scalise,
University of Oregon, United States
Olga Kunina-Habenicht,
Pädagogische Hochschule Karlsruhe,
Germany

***Correspondence:**
Frank J. Blaauw
f.j.blaauw@rug.nl

† These authors have contributed equally to this work

Specialty section:
This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Psychology

Received: 22 November 2018
Accepted: 22 July 2019
Published: 20 September 2019

Citation:
Blaauw FJ, van der Gaag MAE, Snell NR, Emerencia AC, Kunnen ES and de Jonge P (2019) The u-can-act Platform: A Tool to Study Intra-individual Processes of Early School Leaving and Its Prevention Using Multiple Informants. Front. Psychol. 10:1808. doi: 10.3389/fpsyg.2019.01808

We present the u-can-act platform, a tool that we developed to study the individual processes of early school leaving and the preventative actions that mentors take to steer these processes in the right direction. Early school leaving is a significant problem, particularly in vocational education, and can have severe consequences for both the individual and society. However, the prevention of early school leaving is hampered by a mismatch between research and practice: research tends to focus on identifying risk factors using group averages and cross-sectional studies, while practitioners focus on intervening in individual processes. We aim to help solve this mismatch with our project *u-can-act*. In this project we have developed a platform that helps to gain insight into both the individual processes that precede early school leaving as well as the actions that mentors take to prevent it. In this paper we introduce the u-can-act platform, which consists of three technology-based, reusable methodological innovations. Specifically, our innovations concern: (i) an open source web application for longitudinal personalized data-collection, (ii) an automated study protocol that optimizes adherence in a difficult target group (adolescents at risk for early school leaving), and (iii) a technologically assisted coupling between mentor and student that allows us to study dyadic interactions over time. We present performance results of our platform, including participant adherence, the behavior of the questionnaire items over time, and the way that our web application is experienced by the participants. We conclude that our innovative platform is successful in collecting multi-informant time-series data on intervention processes among students in vocational education, both for at-risk students and control students, and for their mentors. Moreover, our platform is suitable for broader applications: it can be used to study any malleable individual process including the efforts of a second individual who aims to influence this process. Because of the unique insights that the u-can-act platform is able to generate, the platform may ultimately contribute to solving the mismatch between research and practice, and to more effective interventions in individual processes.

Keywords: early school leaving, ecological momentary assessments, web application, vocational education, motivation, open source

1. INTRODUCTION

Each year, many adolescents and young adults leave school early[1]. In Europe alone, 5.5 million individuals left school early in 2012 (European Commission, 2013). Early school leaving is a particularly large problem in vocational education and training (VET), where approximately two thirds of all European early school leaving takes place (Cedefop, 2016). This is alarming as early school leaving has severe consequences for both the individual and society as a whole. For example, compared to individuals who obtain a starting qualification, early school leavers have a weaker position on the labor market (e.g., a higher risk of unemployment, lower income, more precarious work conditions) and experience less health, a lower life expectancy, and less life satisfaction (Cedefop, 2016).

Thus it comes at no surprise that on the one hand scientists have spent much effort to investigate the causes and consequences of early school leaving and on the other hand practitioners have spent much effort to try to prevent early school leaving. However, the efforts of the practitioners are not always optimally informed by science. This sub-optimal information may be due to a mismatch in focus: while social scientists have a long tradition of generating knowledge on the *between-individual* level (e.g., finding general trends based on data retrieved from groups), practitioners tend to focus on the *within-individual level* (i.e., individual change processes). This mismatch has two important consequences.

Firstly, our body of between-individual scientific knowledge has facilitated the identification of individuals at risk for early school leaving but has hardly informed prevention strategies. For instance, it has been shown that early school leaving is more likely to occur among males, individuals with a migration background, and individuals with a low social economic status (Rosenthal, 1998). Although this general information is valuable for identifying at-risk individuals, it has little utility to steer interventions of a practitioner, as it is impossible for a practitioner to adjust these factors. Other, more malleable factors have also been demonstrated to be risk factors for dropout, such as problem behavior or negative attitudes toward school (Rumberger and Lim, 2008). Even though a focus on malleable factors is already more useful to the practitioner, merely focusing on malleable factors is still too limited, as reducing risk factors is not the same as promoting graduation and positive youth development (Zaff et al., 2017). In order to perform such promotion, more knowledge is needed on how within-individual processes of positive, malleable factors that are known to promote graduation, such as motivation and engagement (Zaff et al., 2017), can be directly affected by practitioners who work with adolescents.

Secondly, it is fundamentally ill-advisable to use between-individual knowledge to inform within-individual processes. Although research on the between-individual level can provide general information about group characteristics, it provides knowledge that is true on average, but that might not hold true for any individual in specific (e.g., *the non-existent average individual*; Allport, 1937; Blaauw, 2018). Moreover, between-individual knowledge may obfuscate the relations on the individual level, meaning that findings on the between-individual level may not exist on the within-individual level, and can indeed even be opposite (e.g., *Simpson's paradox*; Simpson, 1951; Blyth, 1972, *the ecological fallacy*; Piantadosi et al., 1988, and *non-ergodicity*; Molenaar, 2004; Hamaker, 2012). These problems with translating between-individual findings to within-individual processes are thought to be relevant for the majority of psychological processes (Molenaar, 2004; Kievit et al., 2013). As such, in order to inform practitioners on the individual processes of early school leaving, and how to steer these in the right direction, within-individual research is a necessity.

Fortunately, technological developments have made it increasingly feasible to study within-individual processes. A prominent method to do this is the Ecological Momentary Assessment (EMA) methodology, also known as the experience sampling methodology (ESM), or diary studies (Csikszentmihalyi and Larson, 1987; Shiffman and Stone, 1998). EMA is a methodology widely used in psychopathology research and behavioral research (e.g., Bolger et al., 2003; Trull and Ebner-Priemer, 2009; van der Krieke et al., 2016a). In an EMA study, a participant completes the same questionnaire for a long period of time, possibly multiple times per day, resulting in a large number of measurements of multiple (psychological) variables within one individual. This type of high resolution dataset can provide insight into the processes of the measured variables over time, and the relations between them, within a specific individual. Moreover, the data about an individual can be used to shed light on intra-individual variability, which would be unknown (or assumed non-existent) in a cross-sectional study.

In this paper we present the open source EMA platform of the *u-can-act* research project that we use to study the developmental processes of early school leaving in students, their micro-level interactions with their mentors, and the prevention of early school leaving within individuals. The platform aims to help researchers to effectively study dynamic within-individual processes from multiple informants, even among difficult to reach target groups. It does so by providing an automated way for collecting longitudinal questionnaire data and managing the connections between different informants. The platform can be reused and adapted by other researchers because it is fully open source. The platform is shaped by the aims and theoretical foundations of the u-can-act project, which we present in section 2. The platform encompasses three technological innovations that we present in-depth in section 3, these concern (i) the development of an open source EMA application, (ii) the development of an automated EMA protocol that aims to maximize adherence, and (iii) an innovative coupled multi-informant setup that enables us to investigate dyadic interactions as dynamic processes over time. We collected data among students and their mentors, described in section 4 and use this data to present findings on the performance of the platform. In particular, we focus on its ability to capture

[1] Early school leaving is defined as individuals aged 18–24 who completed at most lower secondary education (International Standard Classification of Education level 2) and who are not involved in further education or training (European Commission, 2013).

within-individual dynamics, the ease of participation for both mentors and students (including early school leavers) and the usability of the platform in section 5. We conclude that our platform is successful in achieving its aim and provide direction for future studies in section 6.

2. THEORETICAL FOUNDATIONS AND AIMS OF U-CAN-ACT

The u-can-act platform and its technological innovations (see also section 3) have their current form because of the aims and the underlying theory that we use in the u-can-act project to study early school leaving and its prevention. U-can-act focuses on (i) malleable, dynamic factors that are relevant to early school leaving (section 2.1) and (ii) dynamic within-individuals processes and dyadic interactions on a weekly, micro-level time-scale (section 2.2). This allows us to clarify processes that precede early school leaving and determine the effects of the mentors' preventative actions on the development of the student, and ultimately, on the students' early school leaving intentions. With this information we aim to inform practitioners on a very practical and detailed level on the actions to take and when to take them, and help policy makers to choose preventative strategies that seem beneficial in reducing early school leaving. We have translated these aims in a theoretical model that reflects our main assumptions (section 2.3). This theoretical model forms the basis of our u-can-act platform.

2.1. A Focus On Malleable Factors

We focus specifically on malleable factors that are expected to vary over time within individuals, and that have the potential to not only prevent early school leaving, but to also promote positive development. A central theory we use for this is the *self-determination* theory. Self-determination theory is an important aspect of the process of early school leaving, while at the same time it is also an important means to promote positive development and intervene in the process of early school leaving (Vallerand et al., 1997; Zaff et al., 2017). The self-determination theory, as proposed and investigated by Deci and Ryan (2012), is primarily a theory of motivation. It postulates the existence of three basic psychological needs, which are autonomy, relatedness, and competence. The fulfillment of basic psychological needs fosters intrinsic motivation, but has recently also been ascribed a broader function: Deci and Ryan (2012) describe that the fulfillment of these needs is "essential for optimal development and functioning" (p. 417). Indeed, need fulfillment has empirically been related to many indicators of well-being and growth, while the frustration of needs is related to illbeing and maladaptive functioning (Vansteenkiste and Ryan, 2013), and of course, early school leaving (e.g., Hardre and Reeve, 2003; Alivernini and Lucidi, 2011).

The interesting characteristic about psychological needs is that they are changeable and can be supported (Hardre and Reeve, 2003; Ntoumanis, 2005; Mouratidis et al., 2011)—thus they form a particularly interesting source of information for practitioners. In fact, in a Dutch study that investigated fifteen early school leaving prevention and intervention projects it was found that the large variety of approaches could be uniformly characterized as aiming to support the autonomy, competence, and relatedness of the students (Heemskerk et al., 2018).

Besides need fulfillment, we focus on two other malleable variables relevant to early school leaving: engagement and expected success. Engagement is an important, malleable factor in the process of early school leaving (Fredricks et al., 2004) and can be defined in several ways (Nielsen, 2016), we chose to focus on two of these. First, behavioral engagement, which is a form of engagement that emphasizes involvement in activities, and is considered crucial in attaining positive academic outcomes and preventing dropout (Fredricks et al., 2004). This is perhaps the most commonly studied form of engagement, but has also been criticized to be one-sided and behavioristic (Nielsen, 2016). Therefore we also study emotional engagement, which has also been found to be an important predictor of early school leaving, besides behavioral engagement (Wang and Fredricks, 2014). In addition to engagement, we focus on the expectations that the students have about the academic success that they will obtain during the school year, as such expectations have also turned out to be malleable yet important predictors of persistence and school success (Zaff et al., 2017).

2.2. A Focus On Individual Processes On a Micro-Level

Much is still unknown about psychological need fulfillment and engagement as part of within-individual, micro-level processes that may change over a short time-span, like weeks or even days. However, some first steps have been made, for example by van der Kaap-Deeder et al. (2017). They found that a sense of autonomy satisfaction or frustration was directly influenced by daily interactions. Moreover they found that the social contexts of these interactions matters, as each of the three social contexts they studied (interactions with mothers, teachers, and siblings), uniquely contributed to whether autonomy satisfaction or frustration is experienced.

Thus experiences in different contexts have the potential to either fulfill or frustrate psychological needs and a within-individual approach is necessary to understand the long-term consequence that this may have for early school leaving. For example, Aelterman et al. (2016) propose that need fulfilling activities result in a pull on the individual, attracting the individual to spend energy on the target activity, while need thwarting activities push the individual away. Extending this hypothesis, we can imagine that in some individual cases need fulfillment may in fact increase the chance of dropout: when individuals spend all their time outside of school because of the need fulfilling context, their engagement with school may decrease and dropout may eventually follow. Such a hypothetical process contradicts the common group-finding that need fulfillment is generally beneficial (Vansteenkiste and Ryan, 2013) and remains unexplored in studies so far because of their inter-individual focus (see also section 1). We can only gain insight into the existence of such hypothetical individual processes by taking a within-individual approach.

Moreover, a micro-level, within-individual approach is necessary in order to learn more about the role that mentors play in influencing students' development and preventing dropout. Individual guidance has proven to be quite effective to prevent early school leaving in many independent intervention and prevention programs (Heemskerk et al., 2018), but much is still unknown about the ingredients of such guidance. Which concrete actions do mentors take in their guidance of adolescents, what goals do they strive for? How effective are they in supporting the basic psychological needs of their students from week to week? Such questions can only be answered by studying the within-individual guidance processes of mentors and the micro-level interactions between students and mentors.

2.3. The Theoretical Foundations of the u-can-act Platform

We placed the malleable factors relevant for early school leaving in a hypothetical model that reflects our within-individual process approach (see **Figure 1**) and have used this model to as the foundation for the u-can-act platform. The interplay between the student and different contexts is at the heart of our theoretical model. Indeed, our within individual approach has led us to hypothesize that the interplay of need fulfillment inside and outside of the school is an important process underlying early school leaving, while it is at the same time a process that a mentor can potentially influence in order to prevent early school leaving. Because we are particularly interested in informing mentors on what they can do to help prevent dropout, the student-mentor interaction is central in our model.

Figure 1 schematically represents the hypothetical model. It reflects the main theoretical assumptions that have driven the development and innovations of our dual informant EMA-platform, and includes the measures that we have employed. These measures cover different aspects of individuals' experience, mental state, and behavior, that are hypothesized to be relevant for the process of early school leaving and interventions in this process. Perhaps the most important assumption that is reflected by this theoretical model, is that students continuously interact with several environments, including a school environment, other environments (non-school, such as the home environment) and their mentor. We included the students' experiences of events and need fulfillment in both school and non-school contexts, as well as experiences of mentor need support and quality of the guidance they receive. We operationalize the students' mental state as emotional engagement, current school success expectations, and well-being. We measure the students' behavior by assessing the amount of time they spent on school activities and how open they have been with their mentor. Similar to the students, mentors have experiences, mental states, and behavior as well, which we operationalized solely with variables relevant to the student-mentor interaction. We assume that the mentors can experience various degrees of satisfaction in their interaction with the student. As a mental state, they can also have various degrees of intuitiveness when performing their actions (as opposed to performing planned actions), and have certain goals they want to achieve. Ideally, their goals are reflected in their actions, but also in their support of students' needs and in their time-investments in the student. This mentor-behavior can be perceived by the student in the quality of the guidance and in the support he or she feels in need fulfillment, with which the student-mentor interaction cycle has come full circle.

To test the relations and processes in our hypothetical model we needed a suitable measurement instrument that met at least three requirements. First and foremost, the instrument needed to repeatedly measure individuals over a period of time in order to gain insight into the within-individual dynamic processes of early school leaving. Secondly, the instrument needed to optimally facilitate easy participation, in order to gather enough data. After all, the processes of motivation that could underlie early school leaving might also influence the motivation of students to partake in this study. Thirdly, the instrument needed to be able to collect measurements for both students and their mentors in order to gain insight into their interaction and into the actions that mentors can take in order to prevent early school leaving. For this, a coupling between the two measurements was necessary. Because there were no applications available that met these requirements, we set out to develop such an application: the u-can-act platform.

3. THE U-CAN-ACT PLATFORM

We developed a platform that allows for studying within-individual processes and dyadic interactions within an intervention setting, from a multi-informant perspective. The platform is rooted in three technological innovations.

The first innovation, and the foundation of our data-collection, is the development of a web application that applies a fully automated method for scheduling, sending invitations, and hosting EMA questionnaires. This free and open source application provides participants with a web interface to fill out weekly questionnaires. Our second innovation is a study protocol that optimizes participant adherence among a difficult target group, which includes an elaborate reward system and messaging that is automatically adapted to the participation behavior of each individual participant. The third innovation is the development of a multi-informant EMA questionnaire that allows us to study the process of early school leaving and the preventative actions in this process from both the student and mentor perspective, where the technology behind our platform manages and deals with the multi-informant aspect of our study by automatically coupling the mentors to their students. We will introduce the three innovations in more detail below.

The three innovations are all integrated in one open source software package, developed by Emerencia et al. (2017) and is freely available at http://u-can-act.com.

3.1. Innovation 1: An Open Source Web-Application

Our first innovation is perhaps most fundamental to our approach: an open source web application that measures the developmental processes of students and their micro-level interactions with their mentors. The application schedules

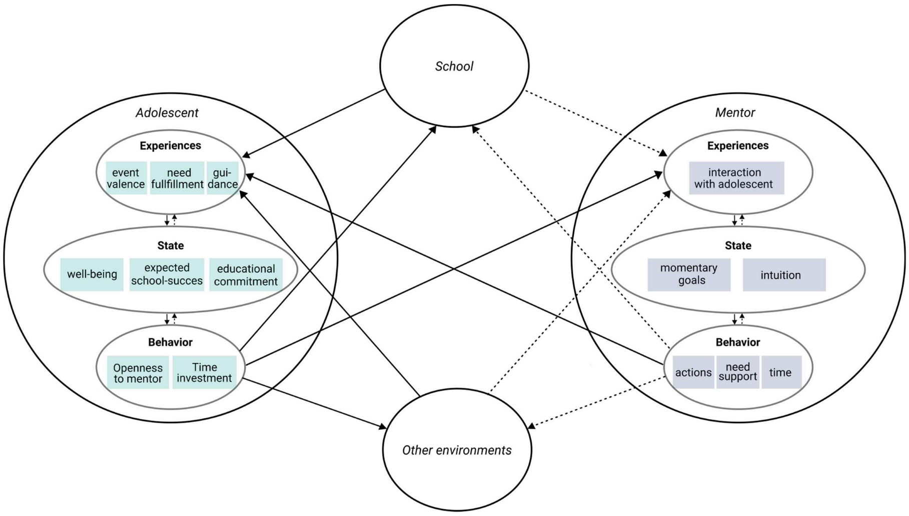


FIGURE 1 | Hypothesized theoretical model of early school leaving that we use in the u-can-act project. In future studies, our platform allows us to study the hypothetical empirical relations that are indicated with a solid line (the relations indicated by the dotted lines cannot be studied in our current set-up).

and sends out questionnaire invites automatically, and stores the data inside two separate and secure databases (one containing personal data and one containing the answers to the questionnaires). Screenshots of the application can be found in the Supplementary Material in **Figure S1**.

A schematic overview of the technological infrastructure of the u-can-act platform is provided in **Figure 2**. The platform serves its content by means of a web application implemented in the Ruby on Rails framework. Ruby on Rails is an open source framework that provides a default structure for web applications[2]. In order for other researchers, schools, and agencies to be free to use and adapt its implementation, we released u-can-act as MIT-licensed[3] open source software on https://u-can-act.com. The implementation of u-can-act builds upon our experience in designing architectures for web-based questionnaire platforms, such as the implementation of the HowNutsAreTheDutch web application (Blaauw and Emerencia, 2015; van der Krieke et al., 2016b).

The collected data is stored into two separate databases: one database that holds the questionnaire data, and one database that contains all personal data. The latter database keeps track of the completed questionnaires by storing a reference to the actual questionnaire data. This separation ensures anonymity in case of a breach in one of these databases. The personal information is stored in a relational SQL database named PostgreSQL. The questionnaire data is stored in a MongoDB NoSQL database. The rationale behind the choice for MongoDB is that it provides a schemaless document storage, which fits well with storing different types of questionnaire data. Finally, we use a third Redis NoSQL database that contains the aggregated / analytical data for caching purposes. Data stored in this database are considered volatile, and mostly used on the researcher dashboard, to provide them with general statistics about the questionnaire completion percentages and rewards collected. Without this cache, these data need to be calculated in real time, which negatively influences the performance of the application.

The traffic to the web application is protected using a 2048 bit RSA (SHA 256 bit) TLS 1.2 Secure Socket Layer (SSL) connection, which ensures private data exchange to and from the u-can-act web application. The interactions with the underlying database infrastructure are protected using SSL as well. Filling out a questionnaire is only possible via the link sent to the participant in a text sent to their phones or in an email message. These links contain a user identifier and a token. The tokens are stored using the Bcrypt encryption standard, which makes it practically impossible to retrieve the clear-text token from its encrypted counterpart.

The platform is built as generic, reusable software, such that other research projects could reuse the platform. Areas in which this could be of interest are, for example, psychiatry (e.g., HowNutsAreTheDutch and Leefplezier; Blaauw et al., 2014a,b; van der Krieke et al., 2016a), general health (Nair et al., 2016), pain monitoring (Stone et al., 2003), substance abuse (Shiffman, 2009), and many other fields that benefit from intra-individual measurements.

The u-can-act application automatically schedules questionnaires and invitations for each participant in the system. During the initial setup phase, the u-can-act application is initialized with a definition of the protocol that contains the collection of measurements, the interval at which invitations should be sent, and the actual questionnaire items that need to be completed. Subsequently all participants can be subscribed to their protocols at any given start and end date. The u-can-act application automatically invites them to complete their questionnaire on a given interval by means of a text message or email.

[2] Website: https://rubyonrails.org

[3] Website: https://opensource.org/licenses/MIT

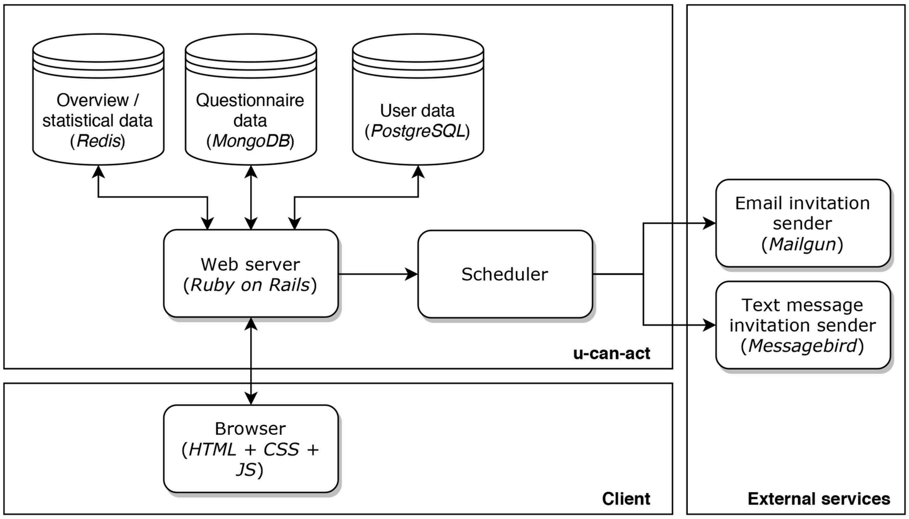

Overview / statistical data (*Redis*)
Questionnaire data (*MongoDB*)
User data (*PostgreSQL*)
Web server (*Ruby on Rails*)
Scheduler
Email invitation sender (*Mailgun*)
Text message invitation sender (*Messagebird*)
u-can-act
Browser (*HTML + CSS + JS*)
Client
External services

FIGURE 2 | Technological infrastructure of the u-can-act web application.

3.2. Innovation 2: Optimizing Adherence and Study Load

Most students included in this study have a high risk of early school leaving, which might also be a risk for their participation behavior in the u-can-act study. Hence, optimizing the study adherence and minimizing the study load has been an important priority for u-can-act. As such, we performed three adherence-optimization steps, which were largely informed by our initial pilot study.

Firstly, we determined an EMA schedule that would work optimally for our sample. From our pilot study, we concluded that the optimal measurement interval is once a week for both the students and the mentors. The main reason for selecting this measurement interval is threefold: (i) this interval coincides well with the frequency of the meetings between student and mentor, (ii) this measurement interval did not significantly reduce the variance in the items compared to more frequent intervals that we also included in the pilot study, and (iii) the evaluation results showed that participants expected this study interval to be most sustainable.

Secondly, in our pilot study we performed interviews, observational studies, and a detailed analysis of each questionnaire question to optimize the users' experience and minimize time-investment while using the application. We optimized the questionnaire questions that scored lowest on understandability and incorporated many qualitative recommendations to increase the usability of the app. This involved, for example, reformulating questions to ask about concrete categories (instead of free text, broader categories, or actions), and providing more information about the meaning, context, and purpose of questions.

Thirdly, and perhaps most importantly, the qualitative data from the pilot study and brainstorm sessions within both our team and one of the involved guidance agencies informed our intrinsic and extrinsic motivational strategies, which we will describe in more detail below.

3.2.1. Fostering Intrinsic Motivation: Personalization

The students receive one SMS text message per week for approximately 35 weeks during the study to inform them that the questionnaire is available for them to fill out. The text messages are framed in a positive way, emphasizing the value of their contribution for their mentor and the research project. The contents of the text messages were dynamically constructed and personalized for each user, taking into account the participation figures (see **Figure S3** in the Supplementary Material for an overview of the invite message and personalization procedure). The rationale behind sending different and personalized text messages was that both the fact that the message text was variable and that it was personalized potentially has a motivating effect for actually filling out the questionnaire (Heerwegh et al., 2005; Muñoz-Leiva et al., 2010).

A second personalization step was performed in the questionnaires themselves. U-can-act uses a system that can automatically tailor questionnaires toward the individual. This means that certain variables are replaced with values relevant to the participant. For example 'your mentor' would be changed to the actual name of the mentor. The options that were available for personalization were: (i) the name of the mentor, (ii) the name of the student, (iii) the gender of the student (different forms), and (iv) the name of the supervisory agency they were affiliated with.

3.2.2. Fostering Extrinsic Motivation: Monetary Rewards for Students

After the EMA study was completed, students received a monetary reward that reflected their amount of completed questionnaires. They received a two Euro reward for each questionnaire they completed. If students completed three questionnaires consecutively, they were awarded a so-called "bonus Euro," which was an additional one Euro reward on top of the two Euro reward. This bonus Euro is an example of *gamification* and aims to motivate the students to complete longer questionnaire series and not leave many gaps, which can be troublesome for certain analyses. The bonus Euro was awarded for each completed questionnaire until one questionnaire was missed, after which the students again needed to complete three consecutive questionnaires. After each completed questionnaire a reward page was displayed to the students. On this page they could see the monetary rewards that they had already earned, the rewards that were still earnable, their progress toward the end-goal (the maximum amount of reward) and their bonus streak. All this was displayed using a playful design, see **Figure S1** in the Supplementary Material for a visualization.

3.3. Innovation 3: Multi-Informant EMA to Study Students, Mentors, and Their Interactions

The u-can-act platform maps out the process of early school leaving and preventative actions in this process from two perspectives: students and mentors. On the one hand, u-can-act collects weekly data about students and their own experiences. On the other hand, the platform takes the perspective of the mentors into account, by asking them to complete questionnaires for each of the students that they supervise. The database is set up in such a way that an automatic coupling is made between each student and their mentor, which enables us to study the interactions between them. Moreover, this coupling helps foster personalization (see also section 3.2), as for example, students see their mentor's name when answering questions about the quality of his or her supervision. We provide more detail on the data collection among students and mentors in sections 4.2 and 4.3.

4. METHODS FOR EVALUATING THE PLATFORM

We collected empirical data among students and mentors during the u-can-act project that we use to evaluate the performance of the u-can-act platform and its three innovations. For this evaluation, we check whether the platform meets three requirements (see also section 2.3) that we believe are essential in order to measure within-individual dynamic processes among adolescents and the interactions with their mentors: (i) dynamicity of the measures, (ii) easy participation, and (iii) good user experience. We describe our data-collection protocol and measurement instruments for both the student as well as the mentor study and give a brief description of our methods for analyzing the performance of the platform.

4.1. Ethics

The u-can-act research protocol was assessed and approved by the ethical committee of the University of Groningen under code 16351-O. All participants provided their informed consent online. No explicit informed consent was collected from the parents/legal guardians of non-adult participants, as all participants were above the age of sixteen.

4.2. Student Study

The first students were enrolled in the student study on November 6, 2017. Students and mentors joined the study on six moments, for an overview see **Figure 3**. The students that participated were all participating in secondary vocational education in three locations spread throughout the Netherlands. The students that participated in this study could be in one of two sub-groups: an *at-risk* subgroup, or a *control* subgroup. The students in the at-risk subgroup were considered to be at risk of early school leaving by their own educational institution, for example because their grades were low, they attended only few classes, experienced stressful situations at home, or showed disruptive behavior in class. Because of this elevated risk, these students were signed up for extra individual guidance. The individual guidance was supplied by mentors from three different supervision agencies (more on this in section 4.3). We approached the students through their mentors: we first asked the mentors to participate, who then asked their students to participate. The students in the control subgroup did not have a mentor, as they were not considered to be at risk for early school leaving and were approached directly.

The student study comprises three main questionnaires: (i) a general assessment, (ii) an EMA questionnaire, and (iii) a post-assessment. The general assessment collected information about the students' demographics and living situation. The EMA questionnaire collected information on variables that could fluctuate over time and are hypothesized to potentially underlie early school leaving (i.e., autonomy, competence, and relatedness). The post-assessment collected information about their current educational situation, such as whether they were still enrolled in their educational track, and whether they intend to complete the track.

4.2.1. Procedure

In order to participate in the study, a student had to be subscribed to the u-can-act platform and provide online informed consent. The control subgroup students were randomly selected from one educational institution in the Northern part of the Netherlands. In collaboration with this educational institution we sampled several students that were considered to be not be at risk of early school leaving, and had not received additional supervision from within their educational institution to help them with school or private problems. If they agreed to participate and accepted the informed consent, they were enrolled in the study.

All students in this study followed the same assessment protocol. Near the start of the EMA study, students were asked to complete a required general assessment questionnaire. Then, for approximately 35 weeks (or until the beginning of the summer holiday period, whichever was shorter), they received a personalized text message each Thursday at noon, in which they were requested to fill out a questionnaire. Each text message contained a link to the u-can-act web application that provided access to the questionnaire they had to fill out. The application automatically sent a reminder text message 8 h later in case a student did not complete the questionnaire before that time. Questionnaires were available for 30 h after the initial invitation. To facilitate early stopping from the study, students were

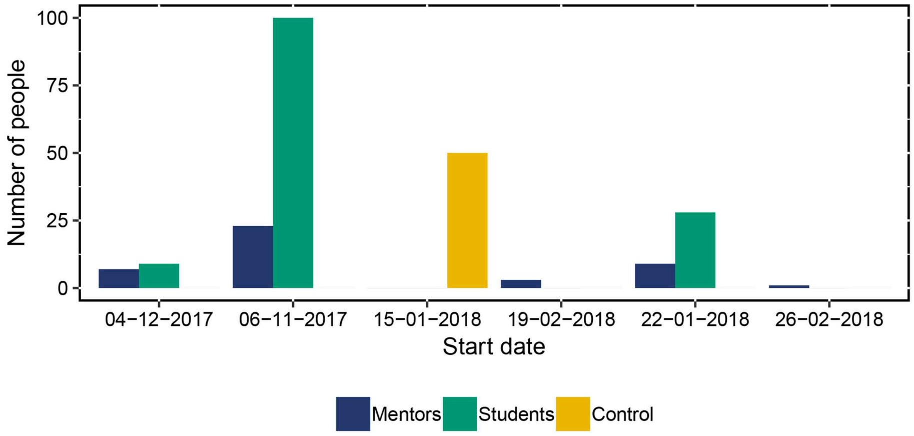


FIGURE 3 | Dates of enrollment of students and mentors.

presented with a button with which they could unsubscribe from the study on June 28, 2018. This button presented them with the question whether their summer holiday had already started, and if it did, that they could end their subscription now, after which they would receive a final, post-assessment questionnaire.

4.2.2. Student Questionnaire Items

The general assessment consisted of nine questions, with which we collect data about (i) birth year, (ii) nationality, (iii) relationship status, (iv) whether or not they had children (and how many), (v) the name of the school they attend, (vi) the type of education they follow, (vii) the level of education, (viii) how many years of education they followed thus far, and (ix) what the students did before starting their current studies. The full list of questions and corresponding answer options is presented in the Supplementary Material in **Table S1**. We collected data about gender during the sign-up process, along with first name, last name, and mobile phone number.

The weekly student EMA questionnaire consisted of twenty-five questions. These questions were in most cases newly created for the purpose of this study, or adapted from previous questionnaires. All questionnaire items are described in **Table S2** in the Supplementary Material. The questionnaire items were selected to assess experienced autonomy, competence and relatedness in three contexts (school, outside-of-school, and mentor relationship); behavioral and emotional engagement with school; school success expectations; evaluations of their mentors' actions; their general level of well-being and the general valence of their school experiences. An interactive example of the web application can be found online[4]. Note that for the control group, all questions related to supervision of a mentor were removed as they were not applicable (questions 18–24).

The visual design of the questionnaire is composed of three different question options: (i) visual analog scales (VAS), (ii) radio buttons, and (iii) checkboxes. Each of the VAS scales provides a continuous value ranging from 0 to 100, and displays a small indicator showing the selected number. The default value of the VAS scale was set to 50 (the center of the scale), and the extremes of the scale had appropriate labels (e.g., "not at all" to "very much," see **Table S2**, "Response range"). The checkboxes and radio buttons were used to create multiple choice questions of which, respectively, multiple or only a single answer could be selected. In some cases, the radio questions had an option which allowed for the input of free text.

The post-assessment questionnaire consisted of at least 11 and at most 14 items (depending on the answers to the questions). The questionnaire focused on (i) whether the student dropped out or not (and when), (ii) the average grade of the student, (iii) if the students dropped out we asked whether they would start a new study and if the students persisted, how certain they are that they will complete this study, (iv) their average grade, (v) the quality of the supervision of the mentor, and (vi) some general questions related to the evaluation of the web application. The full questionnaire is provided in **Table S3** in the Supplementary Material.

4.3. Mentor Study

The mentor study started at the same date as the student study, November 6, 2017 (see **Figure 3** for more information), and consisted of three personal self-report questionnaires: a general assessment, a post-assessment and a series of EMA questionnaires about the students that they supervise. Each mentor completed diary questionnaires about their mentoring of each of their students separately. As such, the mentors essentially participated in several parallel EMA studies, one for each of their students.

4.3.1. Procedure

The enrollment procedure for mentors was similar to the student enrollment procedure, although mentors could only participate whenever the mentor was actually actively involved in the supervision of one or more students. We asked the mentors to provide some general, personal information in a general assessment questionnaire. This general assessment questionnaire consisted of four questions concerning (i) education level, (ii)

[4] Website: https://app.u-can-act.nl/dummy/student

year of birth, (iii) years experience in supervising students, and (iv) nationality. The questionnaire and its items are listed in **Table S4** in the Supplementary Material. Note that the gender for each participant was already known upon sign-up.

Similar to the student study, mentors received a weekly text message on Thursday around noon. In addition to the text message, mentors also received an email. Both the text message and email contained an invitation text and a link to their mentor dashboards, which provided each mentor with an overview of the questionnaires they had completed for the students that week and the adherence of each of their students by means of a heat map. This information could be used by the mentors to intervene if too many measurements were missed by a student. An illustration of this dashboard is provided in **Figure S2** in the Supplementary Material. The mentors did explicitly not have access to the actual questionnaire data provided by the students, in order to provide anonymity to the students.

An interactive version of the mentor dashboard and mentor questionnaire is available online[5]. The system automatically reminded the mentors via e-mail and text message to fill out all the questionnaires if they had not done so 8 h after the initial invite.

At the end of the study, or after the mentors clicked a button telling the system that their holiday had started, a mentor received a post-assessment questionnaire. The post-assessment had a dynamic number of items, depending on the number of students that they supervised. It contained six questions related to their experience with supervising students, general questions related to the web application, and one question for each of the students they supervised, asking whether and by how much the student has improved during the supervision phase. The full list of questionnaire items for the post-assessment questionnaire is provided in **Table S7** in the Supplementary Material.

4.3.2. Mentor Questionnaire Items

The mentor EMA questionnaire was constructed in a bottom-up fashion: we designed the instrument in several brainstorm and focus-group sessions with one of the mentoring initiatives. An important outcome of these sessions was a categorization of the actions and goals that mentors frequently take in their guidance of students. In this way we aimed to measure variables that are highly relevant to the mentoring process. All mentor questionnaire items are listed in **Table S6** in the Supplementary Material.

The mentor questionnaire was different from the student questionnaire in the sense that it was partly dynamic based on the needs of the mentor and could consist of a varying number of questions. By default, the questionnaire contained 24 questions, which could dynamically be extended to a maximum of sixty-nine questions depending on the information a mentor wanted to provide. The dynamic part of this questionnaire resides in its third question, which reads “Add another action (or series of actions).” This question provided the mentors to add up-to ten new action clusters (see **Table S6** in the Supplementary Material) to record actions they had performed for the current student.

4.4. Analysis of the Platform

We show whether our platform indeed captures within-individual dynamics by calculating the root mean squared successive difference (RMSSD) for each of the questions in the separate questionnaires. The RMSSD is a measure of instability, and provides insight into the fluctuation of a variable over time (von Neumann et al., 1941). Fluctuation or variability is important for questions to be meaningful in an EMA (if a question does not fluctuate, there is no value in repeatedly collecting it) and is necessary to capture in order to gain more insight into within-individual processes over time. We calculate the average RMSSD of each of the continuous variables for each participant in separation and then report the average.

Next, we provide insight into the ease of participation by firstly providing a detailed overview of the adherence to the study over time for all the followed subgroups. We zoom into the adherence among students who dropped out of their educational trajectory. Secondly, we show how long it takes to fill in the questionnaire. We have implemented a questionnaire system that records the difference in time between subsequent questions in the questionnaire that allows us to do so. These timings provide a general insight into the ease of answering questions, and into which questions take more time than others and might be candidates for revision in future research.

Finally, we give some preliminary insight into whether our platform is able to take successful measurements among both students and mentors. We do so by reporting on how our participants have experienced the use of our platform using quality indicators from the post-assessment. Specifically, we asked all participants to grade the application on a scale from 1 to 10 (in steps of 0.5). Furthermore, we asked how difficult they found it to keep participating in the study on a scale from 0 to 100, where 0 denotes that it was very difficult to participate and 100 denotes that it was easy to participate.

5. RESULTS

Before we evaluate the platform we first provide some characteristics of our sample. We then evaluate the performance by demonstrating the dynamics of the items, the ease of participation and the user experience of the platform.

5.1. Sample Characteristics

On July 27, 2018 the data collection in the u-can-act project was completed. The data set comprises of a total of 40 mentors from three supervisory agencies that participated in u-can-act, and 181 students, of which 50 are in the control group. We excluded one participant from the dataset because of seemingly unrealistic answer patterns; this individual had left all the sliders on their default value, without manually placing them there. Moreover, we found that individuals with older browsers did not see some questions (hidden questions that were toggled by other questions). This error occurred in 4.9% of the data, which was also excluded. The application was fixed to resolve this error in future studies.

The mean age of the mentors was 33.09 years (median = 28, range 20–49, standard deviation [SD] = 12.62) and 67.44% were

[5] Website: https://app.u-can-act.nl/dummy/mentor

women. The mentors had on average 4.46 years of experience (median = 2, range 0–25, SD = 5.96). Most mentors (95.35%) had the Dutch nationality. 83.33% of the mentors had at least finished intermediate vocational education.

In the at-risk student sample, the mean age was 20.59 years (median = 20, range 16-33, SD = 2.63) and 54.74% were women. The control student sample had similar characteristics and were on average 19.17 years old (median = 19, range 17–25, SD = 1.92) and contained slightly more women (66%). The students started their current study after: high school (at-risk: 39.2%, control: 70.83%), another secondary vocational education trajectory (at-risk: 45.6%, control: 20.83%), working (at-risk: 4.8%, control: 6.25%), or something else (at-risk: 10.4%, control: 2.08%).

There were 17 students who dropped out of their educational trajectory during our study. All of the dropouts were in the at-risk group and none were in the control group. Of these dropouts, 11 left the educational system entirely ("school leavers"), while 6 students had plans to switch to a new educational trajectory ("switchers"). Most of the students, particularly the switchers, dropped out near the end of the academic year in the Netherlands (which coincided with the end of our measurements). For an overview of the dropout moments see **Figure 4**.

5.2. Dynamics of the Questionnaires Items

We calculated RMSSD's to investigate whether our instrument was capable of capturing the dynamics of within individual processes. Over all groups and items, the average RMSSD was 16.22 (median = 13.72, range 7.85–68.47). We also calculated RMSSD's for each of the item separately, these are listed in the tables describing the questionnaire items (**Table S5** for the mentor questionnaire and **Table S2** for the student at-risk and control questionnaire). The RMSSDs indicate that most items showed, on average, reasonable variation, and that none of the questions had drastically more variation than the others. The outlier of question 4 in **Table S5** can be attributed to the fact that this question asked for the time spent on the supervision of a student, and is therefore scaled differently than the other questions, which are ranged 0–100.

5.3. Ease of Participation

Ease of participation was measured by both global adherence numbers (i.e., the number of filled out questionnaires) and the time it took for each questionnaire to be completed.

5.3.1. Adherence to the Study Protocol

Across all agencies, the participants completed a total 6659 assessments On average, each at-risk student that started the diary study[6] completed approximately 68.25% of their possible diary questionnaires. The control group completed approximately 83% of their possible diary questionnaires. The completion rate of the mentors was slightly lower at 52.28%.

The adherence to the study over time is depicted in **Figure 5**. Here, **Figure 5A** shows the general (normalized) adherence to the study over time, in terms of completed questionnaires per group (control students, mentors, and at-risk students). Participation dropped rapidly after week 25, probably because we provided the participants with the option to finish their participation and fill-out the post-assessment questionnaire, as summer vacations started for many of them. In **Figure 5B**, we show the distribution of the percentage completed questionnaires for each of the subgroups. It is interesting to note that most of the at-risk students and the students in the control group completed at least 90% or even 100% of the questionnaires, while only a small part of the mentors showed such consistent adherence.

Additionally, we zoomed in on the adherence behavior of the students who dropped out of their educational trajectory, see also **Figure 6**. We can see two distinct patterns for the two types of dropouts: the school leavers and the switchers. School leavers show a completion percentage over time that is similar to that of the larger at-risk student group, although perhaps surprisingly, they seem to complete more than average questionnaires in the beginning of the year. It is also interesting to see that the majority of school leavers tend to keep participating in our study even after the moment of school dropout. This is different for the switchers, they participate a little less than the larger at-risk group in the beginning of the year, and their participation in our study declines sharply in the 15 weeks before the school dropout moment.

5.3.2. Questionnaire Completion Times

We investigated the time it takes to complete each question, and the questionnaire as a whole. The average completion time for each of the questions for both students and mentors is shown in **Tables S2**, **S5**. **Figure 7A** shows the distribution of completion times as measured over the whole study (i.e., the time it takes to fill out a questionnaire). Very often (in 93.65% of the cases), the questionnaire was completed within 5 min. Since there is a bimodality in the completion times, we calculated the mode for both peaks in the histogram. The first mode is 7 s, which can be explained by a mentor answering that he or she had not seen the student that week. The second mode in the data is 67 s. Mentors and at-risk students had similar completion times, while the control group generally spent less time on the questionnaire. This can be explained by the fact that the control group usually had a questionnaire with fewer questions (control = 19 vs. at-risk = 25, see also section 4.2). **Figure 7B** shows how the time to complete a questionnaire fluctuates over time. There is a steep decline in completion time in the first 2 to 7 weeks, perhaps indicative of a learning curve. After this the completion times become more stable, although they do mildly and gradually decline even further.

5.4. User Experience

As part of the post-questionnaire, we asked both of the student groups and the mentor group to evaluate the u-can-act platform. This questionnaire was completed by 59 at-risk students, 26 control students, and 12 mentors. The control group graded the platform high with an 8 (median = 8.5, range = 5–10, SD = 1.27), as did the at-risk students with a 7.84 (median = 8, range = 3–10, SD = 1.4) and the mentors with a 7.08 (median = 7, range = 5.5–9, SD = 0.93). The control group judged it to be easy to adhere

[6] Thus the participants that provided informed consent.

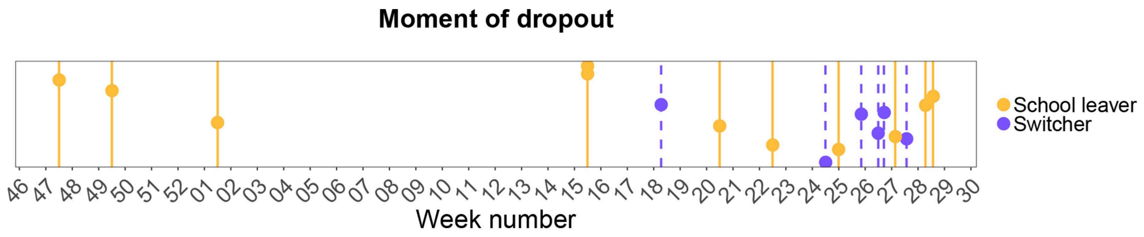


FIGURE 4 | The week of the year in which people dropped out or switched to another study. Each dot is an individual who dropped out.

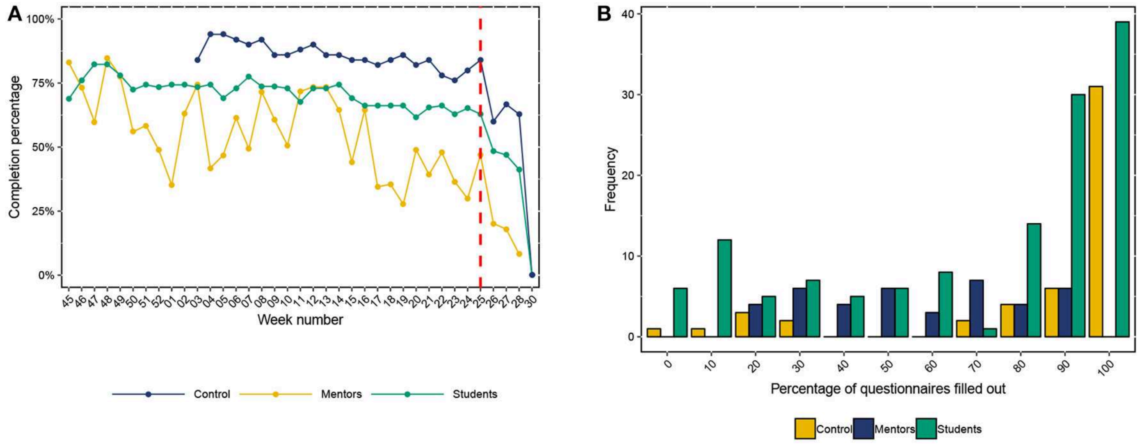


FIGURE 5 | Study adherence for each subgroup. **(A)** Questionnaire completion over time in percentages. The red line indicates when participants could unsubscribe. **(B)** Cross-sectional questionnaire completion in percentages.

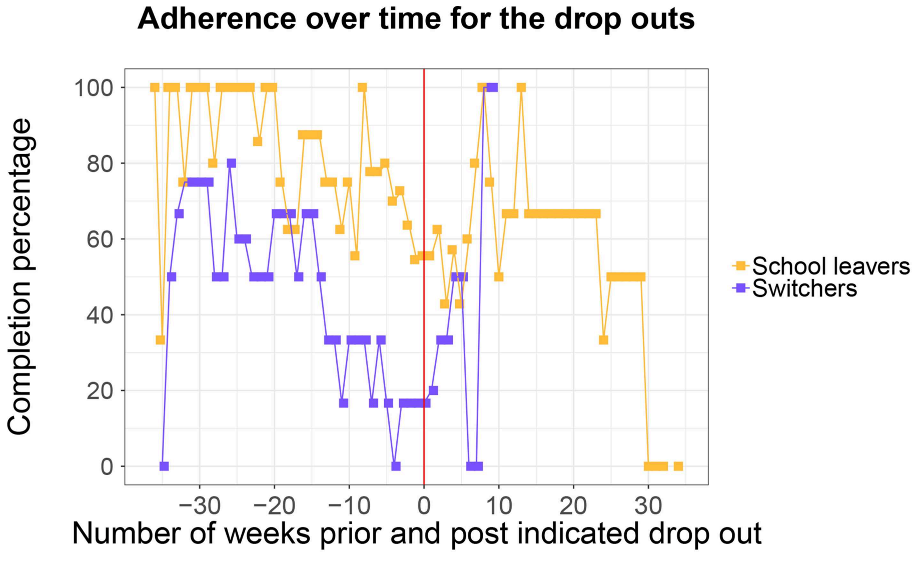


FIGURE 6 | The number of completed questionnaires for the switchers and school leavers. The vertical red line indicates the moment of school dropout.

to the protocol with a mean score of 79.42 (median = 83, range = 0–100, SD = 24.34), as did the at-risk students with a mean score of 73.27 (median = 81, range = 0–100, SD = 28.58) and the mentors found it more difficult than the students with a mean score of 45.67 (median = 45, range = 23–78, SD = 16.93).

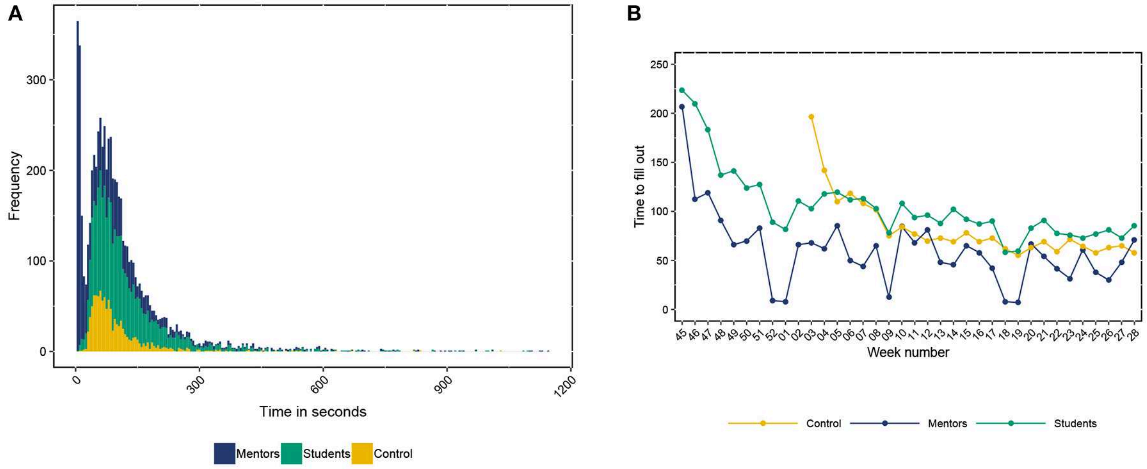


FIGURE 7 | Study adherence for each subgroup. **(A)** The amount of time spent for each questionnaire in the study. **(B)** Median time spent on the questionnaires over time.

6. DISCUSSION

The platform that we have developed within u-can-act seems to be successful in collecting multi-informant and dynamic time-series data on within individual processes among students in vocational education—both regular students and students at risk for early school leaving—and their mentors. This is firstly evidenced by the satisfactory results of the dynamics of our EMA items: their sufficient fluctuation over time, which was on average 16.22 (in terms of RMSSD). The success of our set of innovations is furthermore evidenced by the high participant adherence among a presumably difficult target group: 68.25%. For the control group, the adherence was even higher (83%), signifying the difficulty of the at-risk subgroup (the at-risk students) we are dealing with, while adherence was lowest among the mentors (52.28%). Interestingly, among students who dropped out of their educational trajectory, the school leavers in our sample participated at the same level or even more than the at-risk group, while the switchers participated less. In addition, the questionnaire items took a relatively short time to answer, which was generally less than 5 min for the whole questionnaire. Moreover, the participants were satisfied with the user experience of the app, and indicated that it was easy to adhere to the protocol for an extended period of time, although the mentors experiences more difficulties in this. We will argue that all our (technological) innovations have contributed to these successes.

First of all, the development of the web-based platform and its innovations was essential for participation. This platform resulted in a flexible data-collection application that can be incorporated in students' daily lives by using their own smartphones. The use of a responsive web application had three major advantages: (i) the questionnaires could be filled out on any smart-phone (independent of its operating system), (ii) participants did not need download an app, and (iii) it gave mentors the option to fill out the questionnaires on a PC or tablet. Our platform was designed in such a way that it can automatically remind participants to fill in their questionnaires, to further facilitate easy participation and improve adherence. Another facilitating feature of our platform was the use of identification tokens, which meant that the participants did not need to log in (and thus did not need to remember their credentials).

We hypothesize that the high adherence is also largely influenced by our measurement protocol aimed at maximizing adherence. Because we optimized the usability of the web application by performing an elaborate quantitative and qualitative pilot study, irritations with both the technology and the formulation of the questions were discovered and solved. This led not only to high adherence, but also to a pleasant user experience which we believed helped the users of our platform to participate seriously in our study and improve the validity of their answers. We applied both internal and external motivational strategies to facilitate adherence and generate a pleasant user experience. However, we assumed that it would be unrealistic to solely rely on the intrinsic motivation for adherence of the at-risk students. We mainly dealt with adolescents at risk for early school leaving who, according to literature and our own theoretical model (see section 2) are likely to have trouble with their intrinsic motivational resources for school-related activities (Hardre and Reeve, 2003), which could affect research participation. We fostered intrinsic motivation as much as we considered possible. We used personalized messages in our invitations that were adapted to their participation behavior for example by complementing them on a long streak of filling in the app (fostering the experiences competence) and emphasized our gratefulness for their contribution to both us researchers and their mentor (fostering relatedness). We also used the name of their mentor (agency) in the application to increase the personal relevance. Apart from focusing on intrinsic motivation, we also stimulated their extrinsic motivation, by designing a monetary reward system that uses gamification and playful design elements in the form of bonus-streaks[7]. As has also been found in literature (e.g., Cerasoli et al., 2014), the combination of extrinsic motivational strategies with intrinsic motivational strategies may help foster motivation more than relying solely on intrinsic motivational strategies when it concerns simple tasks (such as filling out a questionnaire).

In the mentor study, we did not have any extrinsic motivational strategies in place, and fostered only intrinsic motivation through the same type of personalization as we did for the students. We made the assumption that their intrinsic motivation would be strong as our research would be of direct importance for the mentoring agency that they were part of, as it would provide them with important information regarding the effectiveness of the actions they take to prevent early school leaving. However, as mentor participation was quite low compared to student participation (mentors: 52.28% vs. at-risk students: 68.25% vs. control students: 83%) and the mentors indicated to experience a medium degree of difficulty in adhering to the protocol, we believe relying solely on their intrinsic motivation was insufficient. We suspect that using extrinsic motivation as a supplement (e.g., a reward system similar to that of the students) may have been helpful and consider this a promising avenue to explore in future research. Furthermore, the relatively low mentor participation may also be improved by re-evaluating the content of the mentor questionnaire. This questionnaire has a qualitative part where mentors fill in the actions that they took in guiding their students, and to next place these actions in suitable categories. This may have been a relatively hard task for some mentors and future studies may look into how this measurement can be made easier.

We believe that the mentor-student connection was essential for study adherence, and was key to get the at-risk students to participate. We approached the at-risk students through their mentor: if students participated, they did so at the behest of their mentor. Furthermore, during the study, the mentors could monitor their students' study adherence, which allowed them to targetedly motivate each student when needed to increase adherence.

6.1. The Innovations Have Produced an Open Source Platform That Collects Multi-Informant Time-Series Data

In order to allow other agencies and researchers to use the u-can-act platform for their own purposes, we released it as open source software. The open source philosophy has several benefits, such as the verifiability of the source code (anyone can inspect the code and verify its logical integrity) and the fact that the software is freely available. The software package includes technical instructions on the use of the software, making it re-usable for interested others. This may be interesting for other researchers or practitioners specifically interested in processes of early school leaving and its prevention. However our platform also serves a broader audience due to its generic implementation. The u-can-act platform can be used by anyone interested to gain insight into within-individual processes and the dyadic interactions or interventions that influence these processes.

Apart from the software being freely available and verifiable, its open source availability could also attract other developers to work on the platform and maintain it past the span of the u-can-act project itself. Maintenance is crucial in a software project in order for it to remain secure and to incorporate updates of external dependencies.

6.2. Limitations

The u-can-act project is an important step to help reduce early school leaving. However, in the present work, we do not yet propose the means to reduce early school leaving. This was not the focus of the present paper. In this paper we aimed to present the platform that we use to collect data about the mentoring process and the process of early school leaving. Our goal with this platform is to generate knowledge that will help reduce early school leaving, but the platform is not by itself meant to directly contribute to this. This may be a direction for future research however, as the current platform can be augmented with a more elaborate dashboard for mentors, on which they could follow the development of their students and adjust their intervention accordingly. The open source nature of our platform allows for such an augmentation to be developed in the future. By presenting our design, platform and initial findings, we have taken a first step in such a direction. And even if this does not happen, we believe that the data that this platform allows us to collect will foster new insights in the individual processes surrounding early school leaving and will eventually help mentors interact with their students in such a way that early school leaving is reduced.

In u-can-act, we focused on a specific subset of the Dutch educational system: vocational education. The reason for this focus is because most early school leaving takes place in this part of the educational system, and moreover, it comprises the largest number of people in the Netherlands. Because of this specific focus, data collected in this study will only be partly generalizable. On the other hand, we argue that generalizing these data might not be useful regardless of the data collected, as in this paper we strongly advocate for a more personalized approach to dealing with early school drop out.

The software is currently in a state where it requires considerable technical expertise to tailor the platform to the needs of a new research group or mentor agency. Setting up the platform requires a few technical steps, such as setting up a server and hosting a database. We have tried to make this as easy as possible with an elaborate manual[8] that is added to the u-can-act web application (Emerencia et al., 2017). Alternatively, the technical implementation and maintenance could be done by a professional company, but then costs would be involved. Thus, even though it is open source, the current platform is like any other questionnaire platform in the sense that it needs expertise to set-up or maintain, or requires costs for external parties to do so. We will leave it up to future researchers to decide.

[7] We did, however, limit the use of gamification elements in order to prevent the application as being seen as "childish."

[8] Available online from: http://u-can-act.com

We are currently working on an interface so that researchers can do as much as possible of the set-up themselves to help overcome this limitation.

6.3. Future Research

In our future work, we aim to provide a solid understanding of early school leaving and methods to prevent this. We aim to test the theoretical model that served as the foundation of the present work. Our studies will include a mapping and profiling of student processes of dropout and persistence, and mentoring processes over time. Moreover, we will investigate the micro-level dynamics within the student, and between the student and the relevant contexts, such as the mentor, school, and non-school context. Our further research will contribute to a better understanding of the process of early school leaving and the prevention of early school leaving.

7. CONCLUSION

The present work set out to describe and evaluate a novel platform, and its technological innovations, that we have developed in our project u-can-act. The platform allows researchers to investigate within-individual processes of early school leaving and interventions in this process. In fact, with some adaptation, the platform can be useful in any situation where insight is needed in within-individual processes and the way that interventions may affect such a process. The rich and unique dataset that we collected with the u-can-act platform allows us to answer many questions related to an individualized perspective on motivation and early school dropout, which were impossible to answer without these data. Moreover, the open source nature of our platform allows other interested agencies or researchers to also collect detailed multi-informant EMA data to better understand within-individual change processes and the effects of interventions.

ETHICS STATEMENT

The u-can-act research protocol was assessed and approved by the ethical committee of the University of Groningen under code 16351-O. All participants provided their informed consent online. No explicit informed consent was collected from the parents/legal guardians of non-adult participants, as all participants were above the age of sixteen.

AUTHOR CONTRIBUTIONS

FB: wrote the initial draft of the manuscript, performed the analysis. MG: wrote the initial draft of the manuscript, did the literature study. NS: performed major revisions, contacted the participating agencies/students, did the literature study. AE: performed major revisions on the manuscript, performed the analysis. EK: performed major revisions on the manuscript, supervised the implementation of the study and the u-can-act platform. PJ: performed major revisions on the manuscript, supervised the implementation of the study and the u-can-act platform.

FUNDING

The u-can-act project was funded by The Netherlands Initiative for Education Research (NRO) grant (no. 405–16–401) received by EK, MG, PJ and Henk Sligte from the Netherlands Organization for Scientific Research (NWO).

ACKNOWLEDGMENTS

The authors want to thank all participants of u-can-act for participating in this research project. The authors also thank Teun Blijlevens of Umanise who improved the user experience of the application, our collaborator the Kohnstamm Institute who helped set up the study and all agencies involved in this study for their enthusiasm and helpful collaboration: Het Buro, Mijn School, and Plusgroep.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01808/full#supplementary-material

REFERENCES

Aelterman, N., Vansteenkiste, M., Soenens, B., and Haerens, L. (2016). A dimensional and person-centered perspective on controlled reasons for non-participation in physical education. *Psychol. Sport Exerc.* 23, 142–154. doi: 10.1016/j.psychsport.2015.12.001

Alivernini, F., and Lucidi, F. (2011). Relationship between social context, self-efficacy, motivation, academic achievement, and intention to drop out of high school: a longitudinal study. *J. Educ. Res.* 104, 241–252. doi: 10.1080/00220671003728062

Allport, G. W. (1937). *Personality: A Psychological Interpretation.* Oxford: Henry Holt and Company.

Blaauw, F. J. (2018). *The non-existent average individual.* (Ph.D. thesis). Groningen: University of Groningen.

Blaauw, F. J., and Emerencia, A. C. (2015). "A service-oriented architecture for web applications in e-mental health: two case studies," in *2015 IEEE 8th International Conference on Service-Oriented Computing and Applications (SOCA)* (Rome: IEEE), 131–138.

Blaauw, F. J., van der Krieke, L., Bos, E. H., Emerencia, A. C., Jeronimus, B. F., Schenk, M., et al. (2014a). "HowNutsAreTheDutch: personalized feedback on a national scale," in *AAAI Fall Symposium on Expanding the Boundaries of Health Informatics Using AI (HIAI'14): Making Personalized and Participatory Medicine A Reality* (Arlington, VA), 6–10.

Blaauw, F. J., van der Krieke, L., de Jonge, P., and Aiello, M. (2014b). Leefplezier: personalized well-being. *IEEE Intell. Informat. Bull.* 15, 28–29.

Blyth, C. R. (1972). On Simpson's paradox and the sure- thing principle. *J. Am. Stat. Assoc.* 67, 364–366. doi: 10.1080/01621459.1972.10482387

Bolger, N., Davis, A., and Rafaeli, E. (2003). Diary methods: capturing life as it is lived. *Annu. Rev. Psychol.* 54, 579–616. doi: 10.1146/annurev.psych.54.101601.145030

Cedefop (2016). "Leaving education early: putting vocational education and training centre stage. Volume I: investigating causes and extent," in *Cedefop Research Paper 57* (Luxembourg: Publications Office).

Cerasoli, C. P., Nicklin, J. M., and Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: a 40-year meta-analysis. *Psychol. Bull.* 140, 980–1008. doi: 10.1037/a0035661

Csikszentmihalyi, M., and Larson, R. (1987). Validity and reliability of the experience-sampling method. *J. Nerv. Mental Dis.* 175, 526–536. doi: 10.1097/00005053-198709000-00004

Deci, E. L., and Ryan, R. M. (2012). "Self-determination theory," in *Handbook of Theories of Social Psychology: Collection: Volumes 1 & 2, 1st Edn.*, chapter 20, eds P. A. M. van Lange, A. W. Kruglanski, and E. T. Higgins (Thousand Oaks, CA: Sage), 416–437.

Emerencia, A. C., Blaauw, F. J., Snell, N. R., Blijlevens, T., Kunnen, E. S., De Jonge, P., et al. (2017). *U-can-act Web-app (Version 1.0)*. [Web application software]. Retrieved from: www.u-can-act.nl.

European Commission (2013). *Reducing Early School Leaving: Key Messages and Policy Support*. Technical report, European Commission Education and Training.

Fredricks, J. A., Blumenfeld, P. C., and Paris, A. H. (2004). School engagement: potential of the concept, state of the evidence. *Rev. Educ. Res.* 74, 59–109. doi: 10.3102/00346543074001059

Hamaker, E. L. (2012). "Why researchers should think 'within-person': a paradigmatic rationale," in *Handbook of Research Methods for Studying Daily Life, Chapter 3*, eds M. R. Mehl and T. S. Conner (New York, NY: Guilford Publications), 43–61.

Hardre, P. L., and Reeve, J. (2003). A motivational model of rural students' intentions to persist in, versus drop out of, high school. *J. Educ. Psychol.* 95, 347–356. doi: 10.1037/0022-0663.95.2.347

Heemskerk, I., van Eck, E., Buisman, M., and Sligte, H. (2018). *Samen op Weg Naar een Startkwalificatie. Evaluatie Van vsv-Projecten in Het Programma Kansen Voor Jongeren van het Oranje Fonds*. Technical Report 984, Kohnstamm Instituut, Amsterdam.

Heerwegh, D., Vanhove, T., Matthijs, K., and Loosveldt, G. (2005). The effect of personalization on response rates and data quality in web surveys. *Int. J. Soc. Res. Methodol.* 8, 85–99. doi: 10.1080/1364557042000 203107

Kievit, R. A., Frankenhuis, W. E., Waldorp, L., and Borsboom, D. (2013). Simpson's paradox in psychological science: a practical guide. *Front. Psychol.* 4:513. doi: 10.3389/fpsyg.2013.00513

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: bringing the person back into scientific psychology, this time forever. *Measurement* 2, 201–218. doi: 10.1207/s15366359mea 0204_1

Mouratidis, A. A., Vansteenkiste, M., Sideridis, G., and Lens, W. (2011). Vitality and interest–enjoyment as a function of class-to-class variation in need-supportive teaching and pupils' autonomous motivation. *J. Educ. Psychol.* 103, 353–366. doi: 10.1037/a0022773

Muñoz-Leiva, F., Sánchez-Fernández, J., Montoro-Ríos, F., and Ibáñez-Zapata, J. Á. (2010). Improving the response rate and quality in Web-based surveys through the personalization and frequency of reminder mailings. *Qual. Quant.* 44, 1037–1052. doi: 10.1007/s11135-009-9256-5

Nair, S., Kheirkhahan, M., Davoudi, A., Rashidi, P., Wanigatunga, A. A., Corbett, D. B., et al. (2016). "ROAMM: a software infrastructure for real-time monitoring of personal health," in *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services, Healthcom 2016* (Munich).

Nielsen, K. (2016). Engagement, conduct of life and dropouts in the Danish vocational education and training (VET) system. *J. Vocat. Educ. Train.* 68, 198–213. doi: 10.1080/13636820.2015.1133694

Ntoumanis, N. (2005). A prospective study of participation in optional school physical education using a self-determination theory framework. *J. Educ. Psychol.* 97, 444–453. doi: 10.1037/0022-0663.97.3.444

Piantadosi, S., Byar, D. P., and Green, S. B. (1988). The ecological fallacy. *Am. J. Epidemiol.* 127, 893–904. doi: 10.1093/oxfordjournals.aje.a114892

Rosenthal, B. S. (1998). Non-school correlates of dropout: an integrative review of the literature. *Child. Youth Serv. Rev.* 20, 413–433. doi: 10.1016/S0190-7409(98)00015-2

Rumberger, R. W., and Lim, S. A. (2008). *Why Students Drop Out of School: A Review of 25 Years of Research*. California {Dropout} {Research} {Project} {Report} #15 October, University of California.

Shiffman, S. (2009). Ecological momentary assessment (EMA) in studies of substance use. *Psychol. Assess.* 21, 486–497. doi: 10.1037/a0017074

Shiffman, S., and Stone, A. A. (1998). "Ecological momentary assessment: a new tool for behavioral medicine research," in *Technology and methods in behavioral medicine, 1st Edn.*, Chapter 7, eds D. S. Krantz and A. Baum (Mahwah, NJ: Lawrence Erlbaum Associates), 117–131.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *J. R. Stat. Soc. Ser. B* 13, 238–241. doi: 10.1111/j.2517-6161.1951.tb00088.x

Stone, A. A., Broderick, J. E., Schwartz, J. E., Shiffman, S., Litcher-Kelly, L., and Calvanese, P. (2003). Intensive momentary reporting of pain with an electronic diary: reactivity, compliance, and patient satisfaction. *Pain* 104, 343–351. doi: 10.1016/S0304-3959(03)00040-X

Trull, T. J., and Ebner-Priemer, U. W. (2009). Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: introduction to the special section. *Psychol. Assess.* 21, 457–462. doi: 10.1037/a0017653

Vallerand, R. J., Social, C., Fortier, M. S., Elliott, A., Blais, M., Vallerand, R. J., et al. (1997). Self-determination and persistence in a real-life setting " toward a motivational model of high school dropout. *J. Pers. Soc. Psychol.* 72, 1161–1176. doi: 10.1037/0022-3514.72.5.1161

van der Kaap-Deeder, J., Vansteenkiste, M., Soenens, B., and Mabbe, E. (2017). Children's daily well-being: the role of mothers', teachers', and siblings' autonomy support and psychological control. *Dev. Psychol.* 53, 237–251. doi: 10.1037/dev0000218

van der Krieke, L., Blaauw, F. J., Emerencia, A. C., Schenk, H. M., Slaets, J. P. J., Bos, E. H., et al. (2016a). Temporal dynamics of health and well-being. *Psychosomat. Med.* 79:1. doi: 10.1097/PSY.0000000000000378

van der Krieke, L., Jeronimus, B. F., Blaauw, F. J., Wanders, R. B. K., Emerencia, A. C., Schenk, H. M., et al. (2016b). HowNutsAreTheDutch (HoeGekIsNL): a crowdsourcing study of mental symptoms and strengths. *Int. J. Methods Psychiatr. Res.* 25, 123–144. doi: 10.1002/mpr.1495

Vansteenkiste, M., and Ryan, R. M. (2013). On psychological growth and vulnerability: {basic} psychological need satisfaction and need frustration as a unifying principle. *J. Psychother. Integr.* 23, 263–280. doi: 10.1037/a0032359

von Neumann, J., Kent, R. H., Bellinson, H. R., and Hart, B. I. (1941). The mean square successive difference. *Ann. Math. Stat.* 12, 153–162. doi: 10.1214/aoms/1177731746

Wang, M. T., and Fredricks, J. A. (2014). The reciprocal links between school engagement, youth problem behaviors, and school dropout during adolescence. *Child Dev.* 85, 722–737. doi: 10.1111/cdev.12138

Zaff, J. F., Donlan, A., Gunning, A., Anderson, S. E., McDermott, E., and Sedaca, M. (2017). Factors that promote high school graduation: {a} review of the literature. *Educ. Psychol. Rev.* 29, 447–476. doi: 10.1007/s10648-016-9363-5

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

PERSPECTIVE
published: 25 September 2019
doi: 10.3389/fpsyg.2019.02168

What Technology Can and Cannot Do to Support Assessment of Non-cognitive Skills

Vanessa R. Simmering, Lu Ou and Maria Bolsinova*

ACTNext by ACT, Inc., Iowa City, IA, United States

Advances in technology hold great promise for expanding what assessments may achieve across domains. We focus on non-cognitive skills as our domain, but lessons can be extended to other domains for both the advantages and drawbacks of new technological approaches for different types of assessments. We first briefly review the limitations of traditional assessments of non-cognitive skills. Next, we discuss specific examples of technological advances, considering whether and how they can address such limitations, followed by remaining and new challenges introduced by incorporating technology into non-cognitive assessments. We conclude by noting that technology will not always improve assessments over traditional methods and that careful consideration must be given to the advantages and limitations of each type of assessment relative to the goals and needs of the assessor. The domain of non-cognitive assessments in particular remains limited by lack of agreement and clarity on some constructs and their relations to observable behavior (e.g., self-control versus -regulation versus -discipline), and until these theoretical limitations must be overcome to realize the full benefit of incorporating technology into assessments.

Keywords: non-cognitive, competencies, assessment, construct validity, technological advances, theoretical limitations

OPEN ACCESS

Edited by:
Frank Goldhammer,
German Institute for International Educational Research (LG), Germany

Reviewed by:
Oliver Luedtke,
University of Kiel, Germany
René T. Proyer,
Martin Luther University of Halle-Wittenberg, Germany

****Correspondence:***
Vanessa R. Simmering
vanessa.simmering@act.org

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *07 December 2018*
Accepted: *09 September 2019*
Published: *25 September 2019*

Citation:
Simmering VR, Ou L and Bolsinova M (2019) What Technology Can and Cannot Do to Support Assessment of Non-cognitive Skills. Front. Psychol. 10:2168. doi: 10.3389/fpsyg.2019.02168

INTRODUCTION

Non-cognitive skills have been increasingly recognized as important contributors to education and workplace success (Levin, 2013). These skills include a wide range of competencies, such as perseverance, collaboration, emotional intelligence, and self-regulation; **Table 1** list those included in a recent systematic review (Smithers et al., 2018). There is some disagreement on how to define and delineate them, including whether such attributes are fixed traits or malleable skills (for discussion, see Lipnevich et al., 2013; Duckworth and Yeager, 2015; Smithers et al., 2018; Simmering et al., 2019). Although these are important theoretical issues that will inform assessment development, they are beyond the scope of the current paper. Rather, we discuss how advances in technology may change non-cognitive assessments. We aim to provide a high-level overview of advantages gained through technology, along with new and remaining challenges that must be addressed. We focus on non-cognitive skills because many are more contextual and dynamic than academic skills (e.g., delay of gratification, emotional reactivity). Before considering technological advances, we first briefly review the limitations of traditional non-cognitive assessments.

TABLE 1 | Non-cognitive skills included in Smithers et al. (2018) systematic review and meta-analysis.

High-level descriptors
Character skills
Executive functions
Personality traits
Socio-emotional skills
Soft skills
Specific capabilities
Attention
Cognitive flexibility/control
Conscientiousness
Delay of gratification
Effortful control/self-control/regulation
Emotional stability/reactivity/regulation
Impulsivity
Inhibitory control
Locus of control
Motivation
Perseverance/persistence
Responsibility
Self-esteem
Sociability

Smithers et al. did not differentiate terms as high-level versus specific; this has been added to acknowledge the multidimensional nature of the high-level constructs, though we recognize that some specific capabilities may also be multidimensional. We also group terms we viewed as synonymous within specific capabilities, although these views are not universal in the broader literature.

COMMON LIMITATIONS IN ASSESSMENTS OF NON-COGNITIVE SKILLS

Duckworth and Yeager (2015) reviewed concerns with measurement of non-cognitive skills, outlining limitations of two types of assessments, questionnaires, and performance tasks, using the construct self-control for illustration (see Simmering et al., 2019, for related discussion). Questionnaires can be administered to any informant but most commonly use parent- and teacher-report for children and self-report for adolescents and adults. Questionnaires may ask about a subject's behavior in general, in a specified period (e.g., at this moment, in the past week, month, or year), or in a hypothetical situation [as in situational judgment tests (SJTs)]. Responses may be ratings of frequency (e.g., "almost never" ranging to "almost always"), how well a description fits the subject (e.g., more or less true or like the individual), or choices of specific behaviors in SJTs. The limitations Duckworth and Yeager described were misinterpretation of items, lack of insight or information, insensitivity at different time scales, and reference or social desirability bias. Simmering et al. (2019) also noted context insensitivity as a limitation, as behaviors may occur in some contexts but not others that are not differentiated by questionnaires (e.g., perseverance in school work versus hobbies, or different academic subjects). Furthermore, some studies suggest that self-reports in response to hypothetical situations diverge from actual behavior in analogous experiences (Woodzicka and LaFrance, 2001; Bostyn et al., 2018). Limitations of questionnaires have been extensively studied (e.g., Furnham, 1986), with numerous remedies developed (e.g., Kronsik and Presser, 2009).

An alternative approach is to observe behavior directly rather than eliciting informants' reflection and interpretation. Performance tasks are designed to compel behavior in relevant contexts, with the advantage of creating controlled situations in which all subjects are observed (for discussion, see Cronbach, 1970). For example, objective personality tests assess personality traits through behavioral indicators from performance tasks rather than self-reports (Ortner and Schmitt, 2014). Although performance tasks offer advantages over questionnaires – avoiding subjective judgments by informants, less opportunity for social desirability, reference, and acquiescence biases, more temporal sensitivity – they have serious limitations (see Duckworth and Yeager, 2015; Simmering et al., 2019, for further discussion). For example, lab-based performance tasks such as the Balloon Analogue Risk Task (Lejuez et al., 2002) typically assess single constructs (i.e., risk-taking) and lack diversity needed to form a complete personality profile. Performance tasks are generally designed to elicit one "right" behavior and may conflate "wrong" behaviors that reflect different underlying causes (e.g., Saxler, 2016). Participants' behavior may reflect factors beyond the intended construct, such as compliance with authority of comprehension of instructions. This is a particular concern when participants' prior experiences differ substantially from those designing, administering, and interpreting the tasks; behavior considered maladaptive in the task may be more appropriate to participants' experience. Furthermore, task artificiality could create inauthentic motivations and constraints, leading to unnatural behaviors. Tasks with scenarios created in real time can also lead to error in task implementation, recording of behavior, or participant responses.

To overcome these types of limitations, Duckworth and Yeager (2015) recommended using multiple measures suited to the assessor's goals while acknowledging and accounting for the limitations of each. They also noted that further innovation in assessment could avoid some limitations, with specific examples including incorporation of technology. In the next section, we review technological advances in non-cognitive assessments and the advantages they offer.

ADVANTAGES OF TECHNOLOGY-ENHANCED ASSESSMENTS

Technology allows new and expanded ways to collect data and present content. Computerizing assessments has become more common as access to technology has increased, but these implementations often merely reconfigure prior assessments to be presented on a screen without further adaptation. We focus on more substantive changes that expand the scope of the types of measurements and content included in non-cognitive assessments.

First, technology allows for real-time collection of multiple types of data, including self-reports, physiological data, and observed behavior. Traditionally, assessments are presented once or a few times at widely spaced intervals. Continuous, unobtrusive data collection is now possible through devices such as smartphones or fitness trackers. For example, Wang et al. (2014) combined multiple data sources from automated sensors on a smartphone (i.e., accelerometer, microphone, light sensing, global positioning, Bluetooth) with self-report sampling to evaluate how college students' daily activity related to their mental well-being (i.e., depression, stress, loneliness) and academic performance. Sensor data correlated moderately with these outcomes, as well as students' self-reports. These data were then used to infer students' studying and social behavior to predict their GPA (Wang et al., 2015), indicating how sensor data could be used instead of self-reports. Automated sensors are not only less obtrusive to participants but can also provide a more temporally complete record, which avoids relying on narrow sampling and extrapolation to track change over time (c.f., Adolph et al., 2008). Such temporal detail is necessary to evaluate dynamic non-cognitive skills, such as self-regulation.

Second, ecologically valid methods allow data collection directly from relevant contexts, avoiding the need for retrospection or generalizations in questionnaires, imagined experiences in SJTs, or contrived scenarios in a lab (see Stone and Shiffman, 1994, for related discussion). Experience sampling methods, such as ecological momentary assessments and daily diaries, ask participants report thoughts, feelings, behaviors, and environment at regular intervals over time or around target events. They have been widely used to track emotions in natural contexts, allowing assessment of emotion regulation (Silk et al., 2003; Tan et al., 2012). When contextual variation is also recorded, these assessments can tally how frequently a subject encounters specific contexts and whether behavior varies across those contexts.

Third, some devices allow data collection not attainable without technology. For example, during computerized activities, participants' eye movements can be continuously recorded using eye-trackers, and mouse movements or touchscreen selections can be collected using specialized software. Such data were inaccessible before technological solutions were developed, and they provide the opportunity for more holistic analysis of behavior. Assessments that provide these and other types of process data during participation, such as item-level response latencies (e.g., Ranger and Ortner, 2011), allow researchers to use more than just final responses to improve measurement. For example, pupillometry and reaction times can differentiate whether participants were controlling attention proactively (i.e., mentally preparing for target actions) versus reactively (i.e., adjusting action following external signals) even when target actions (i.e., identifying a stimulus sequence) did not differ (Chatham et al., 2009). Log files of online game-based assessments include time and event information that can be used to track participants' collaboration during the game (Hao et al., 2016; Hao and Mislevy, 2018). Process data may provide insight into responses that would not be possible without technology, and analyzing such data can support assessment validation (Lee et al., 2019).

Beyond data collection, technology enables presentation of content in ways not possible with traditional assessments. Computerized adaptive testing draws items from a large pool of items with varying difficulty to present them adaptively based on test-takers' previous responses and estimated ability (Segall, 2005). This allows more sensitivity to student ability levels and reduces the influence of small mistakes and lucky guesses on the final estimated ability. While computerized adaptive testing is most often used to measure cognitive abilities, it can also improve the measurement of other constructs, like personality (Makransky et al., 2013) and mental health (Becker et al., 2008; Stochl et al., 2016). Because adaptivity is an important facet of non-cognitive skills, test design and administration organizations such as the National Center for Education Statistics recommend adaptive tests in collaborative problem solving and other future assessments (Fiore et al., 2017).

Beyond contingent item presentation, interventions can also be integrated into computerized assessments. Based on assessment results, personalized feedback and recommended learning materials can be provided to respondents to improve individual development. Such systems have gained popularity in assessments of cognitive skills (e.g., Klinkenberg et al., 2011) but can also support non-cognitive skills. For example, Hutt et al. (2017) developed an eye-tracking application to monitor students' mind-wandering in real time during a computerized learning task. When mind-wandering is detected, the application intervenes to repeat the recent material, redirect the student's attention, or ask a question to allow self-reflection in the student. Although the goal was to improve students' learning of the material, feedback on the frequency of mind-wandering could also teach students to monitor and regulate their mental engagement.

The nature of the material going into assessment items can also be expanded by technology. Rather than presenting text questionnaires, researchers can create multi-modal vignettes to present scenarios like SJTs. Audio-visual presentation is preferable to text for students with limited reading comprehension and can increase the validity for such groups (e.g., Chan and Schmitt, 1997). Through interactive technology like digital games and virtual or augmented reality, more complex content can be created to simulate real-life contexts that may be difficult to observe naturally. These environments can include "stealth" assessments in which students' capabilities are evaluated without explicit queries. For example, in a role-playing game comprising quests that require creative problem solving, players' actions may be scored for evidence of both cognitive (e.g., reading comprehension) and non-cognitive (e.g., persistence) competencies (Shute, 2011). Embedding target constructs in naturalistic interactions allows participants to respond with authentic behaviors rather than reporting imagined behavior in response to a hypothetical scenario. This can increase motivation and engagement when properly designed (Moreno-Ger et al., 2008), which in turn could reduce measurement error.

Technological advances can also facilitate generation of new content with reduced human effort, a vital feature for delivering assessments at scale. Machine learning and artificial intelligence have been developed for generating traditional assessment content (i.e., item stems and response options), although much work remains to achieve wide adoption (Gierl et al., 2012). One potential advantage to automated content generation, beyond the efficiency, is the expanded ability to personalize material for students. For example, research on motivation and engagement suggests that integrating students' social and cultural identities into instructional and assessment design can improve outcomes for students from marginalized groups (Haslam, 2017). More work is needed to identify the best ways to design non-cognitive assessments to align with students' identities, but technology provides a promising avenue to realize this level of personalization.

CHALLENGES IN ADOPTING TECHNOLOGY-ENHANCED ASSESSMENTS

Technology-enhanced assessments are not without challenges and limitations. First, construct validity remains a significant concern, and adapting previous assessments to incorporate new technology may affect validity positively or negatively. As noted above, video vignettes in SJTs increased validity by decreasing the influence of reading comprehension (Chan and Schmitt, 1997). Conversely, more complex scenarios could introduce variation in interpretations or decision processes by participants. Such complexity likely reflects real-life contexts more closely but introduces challenges for standardization, especially when content presentation is contingent on participant performance. Standardized items and tasks, as well as scoring rubrics, for virtual performance assessments must be developed and validated in pilot studies (Hao et al., 2017).

The collection of more extensive, ecologically valid, and objective measures of behavior, whether during natural experience or games and simulations, still requires interpretation of how behaviors relate to underlying constructs (an important facet of construct validity; Borsboom et al., 2004). For example, although Hutt et al. (2017) related pupillometry and saccade duration to mind-wandering, these behaviors could be driven by external factors rather than internal processes. Similarly, data from automated sensors (as in Wang et al., 2014) cannot directly address whether variation in recorded activities reflects internal differences (i.e., participants' self-regulation abilities) versus external forces. It is also possible that behaviors measured in these ways are not representative: knowing one is being observed in daily life may lead to atypical behavior, especially when a device is first introduced (c.f., Alvero and Austin, 2004), or participants may be more willing to act "out of character" in a simulation.

Second, one must consider both ethical issues shared with traditional assessments (e.g., how data will be stored, used, and potentially shared; proper training for those administering and interpreting assessments) and new issues that arise with technology. Technological requirements can contribute to inequity, as not all communities have access to necessary infrastructure (e.g., internet bandwidth, devices meeting specifications) or funding to adopt high-tech assessments, and participants may be unaccustomed to using technology. Automated or continuous recordings may invade the privacy of participants or non-participants who have not consented to have their data collected (e.g., conversation partners in audio recordings); although these concerns would be addressed through human subjects protections for research, such protections do not extend to assessments in non-research settings. Ethical concerns for developing technological assessments are conceptually similar to traditional assessments but may be practically different. For example, machine learning algorithms may be biased due to the training sets used to develop them (Springer et al., 2018) similar to how questionnaires may be biased by validation with unrepresentative samples (Clark and Watson, 2019).

Third, collection of more varied and continuous data introduces challenges in compliance and data management. Participants may find continuous or frequent sampling intrusive and therefore be less willing to complete an assessment. Imperfections in devices and software can lead to lost data, with some sources of loss relating to constructs of interest (e.g., losing track of eye gaze if posture changes as interest wanes). The multitude of possible reasons underlying data loss across different types of sensors and devices, combined with reasons shared with traditional assessments (e.g., selectively omitting responses, attrition), makes addressing missing data both practically and theoretically complex.

How we make use of more and different types of data across sources also presents new challenges. Connecting multiple assessments to the same individual profile requires complex data management solutions to ensure both privacy for individuals and accessibility for those using assessment results. If multiple sources are used simultaneously in real time, the data streams must be synchronized and at compatible granularity. Intensive longitudinal datasets require developing identifiable statistical models that can accommodate irregularly spaced, high-dimensional, noisy, dynamic data, as well as related robust and efficient computing software to make use of them (Chow et al., 2018).

Lastly, there can be a strong temptation to apply new technology to assessment as it becomes available without fully evaluating the potential costs and benefits of its adoption. It is important not to let technological capabilities be the driving factors in assessment development but rather to focus on the need the assessment is serving and whether that need can be better met by technology. New technological applications must be carefully designed and validated even when they seem to be only a minor change from previous methods. For example, moving from text to audio-visual presentation of SJTs introduces decisions for how each character looks and sounds. Participants may interpret or respond to characters' behavior differently based on demographic features (c.f., Renno and Shutts, 2015) or voice intonation, which can unintentionally alter the content from the text version. Each new development will bring in new considerations for how the method reaches assessment goals.

CONCLUSION

Advances in technology have expanded the horizons of what types of assessments are possible and achievable. These expansions can contribute to our understanding of non-cognitive capabilities as well as traditional academic content. The advantages of technology-enhanced assessments include how and what data can be collected, as well as the content that can be presented.

With these advantages come some new challenges in the implementation and analysis of assessments, as well as the familiar challenges of construct and predictive validity that all assessments must address. Whether technology can improve an assessment will depend on details of the construct, the target group, the aims of the assessment, and the desired implementation. Assessment methods should be tailored to the specific conditions at hand. In the context of non-cognitive assessments in particular, more work is needed to arrive at well-defined constructs with clear connections to behavior as we also work to capitalize on the advantages technology offers.

AUTHOR CONTRIBUTIONS

VS conceptualized the topic and all three authors contributed equally to development of the ideas. VS drafted the manuscript, then LO and MB provided critical additions and revisions.

ACKNOWLEDGMENTS

The authors would like to thank Chelsea Andrews for help finding and discussing relevant articles for background research.

REFERENCES

Adolph, K. E., Robinson, S. R., Young, J. W., and Gill-Alvarez, F. (2008). What is the shape of developmental change? *Psychol. Rev.* 115, 527–543. doi: 10.1037/0033-295X.115.3.527

Alvero, A. M., and Austin, J. (2004). The effects of conducting behavioral observations on the behavior of the observer. *J. Appl. Behav. Anal.* 37, 457–468. doi: 10.1901/jaba.2004.37-457

Becker, J., Fliege, H., Kocalevent, R.-D., Bjorner, J. B., Rose, M., Walter, O. B., et al. (2008). Functioning and validity of a computerized adaptive test to measure anxiety (A-CAT). *Depress. Anxiety* 25, E182–E194. doi: 10.1002/da.20482

Borsboom, D., Mellenbergh, G. J., and van Heerden, J. (2004). The concept of validity. *Psychol. Rev.* 111, 1061–1071. doi: 10.1037/0033-295X.111.4.1061

Bostyn, D. H., Sevenhant, S., and Roets, A. (2018). Of mice, men, and trolleys: hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychol. Sci.* 29, 1084–1093. doi: 10.1177/0956797617752640

Chan, D., and Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: subgroup differences in test performance and face validity perceptions. *J. Appl. Psychol.* 82, 143–159. doi: 10.1037/0021-9010.82.1.143

Chatham, C. H., Frank, M. J., and Munakata, Y. (2009). Pupillometric and behavioral markers of a developmental shift in the temporal dynamics of cognitive control. *Proc. Natl. Acad. Sci. USA* 106, 5529–5533. doi: 10.1073/pnas.0810002106

Chow, S.-M., Ou, L., Ciptadi, A., Prince, E. B., You, D., Hunter, M. D., et al. (2018). Representing sudden shifts in intensive dyadic interaction data using differential equation models with regime switching. *Psychometrika* 83, 476–510. doi: 10.1007/s11336-018-9605-1

Clark, L. A., and Watson, D. (2019). Constructing validity: new developments in creating objective measuring instruments. *Psychol. Assess.*, No Pagination Specified-No Pagination Specified. doi: 10.1037/pas0000626 [Epub ahead of print].

Cronbach, L. J. (1970). "Performance tasks of personality" in *Essentials of psychological testing.* 3rd Edn, (New York, NY: Harper & Row), 608–666.

Duckworth, A. L., and Yeager, D. S. (2015). Measurement matters: assessing personal qualities other than cognitive ability for educational purposes. *Educ. Res.* 44, 237–251. doi: 10.3102/0013189X15584327

Fiore, S. M., Graesser, A., Greiff, S., Griffin, P., Gong, B., Kyllonen, P., et al. (2017). *Collaborative problem solving: Considerations for the national assessment of educational progress*: Washington, DC: NCES.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personal. Individ. Differ.* 7, 385–400.

Gierl, M. J., Lai, H., and Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items: Automatic generation of test items. *Med. Educ.* 46, 757–765. doi: 10.1111/j.1365-2923.2012.04289.x

Hao, J., Liu, L., von Davier, A. A., and Kyllonen, P. C. (2017). "Initial steps towards a standardized assessment for collaborative problem solving (CPS): practical challenges and strategies" in *Innovative assessment of collaboration.* eds. A. A. von Davier, M. Zhu, and P. C. Kyllonen (New York, NY: Springer), 135–156.

Hao, J., and Mislevy, R. J. (2018). The evidence trace file: a data structure for virtual performance assessments informed by data analytics and evidence-centered design. *ETS Res. Rep. Ser.* 2018, 1–16. doi: 10.1002/ets2.12215

Hao, J., Smith, L., Mislevy, R., von Davier, A., and Bauer, M. (2016). Taming log files from game/simulation-based assessments: data models and data analysis tools. *ETS Res. Rep. Ser.* 2016, 1–17. doi: 10.1002/ets2.12096

Haslam, S. A. (2017). "The social identity approach to education and learning: identification, ideation, interaction, influence and ideology" in *Self and Social Identity in Educational Contexts*. (New York, NY: Routledge), 33–66.

Hutt, S., Mills, C., Bosch, N., Krasich, K., Brockmole, J., and D'Mello, S. (2017). "Out of the fr-eye-ing pan: towards gaze-based models of attention during learning with technology in the classroom" in *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (Bratislava, Slovakia: ACM), 94–103.

Klinkenberg, S., Straatemeier, M., and van der Maas, H. L. J. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Comput. Educ.* 57, 1813–1824. doi: 10.1016/j.compedu.2011.02.003

Kronsik, J. A., and Presser, S. (2009). "Question and questionnaire design" in *Handbook of survey research*. eds. J. D. Wright, and P. V. Marsden (San Diego: Elsevier).

Lee, Y.-H., Hao, J., Man, K., and Ou, L. (2019). How do test takers interact with simulation-based tasks? A response-time perspective. *Front. Psychol.* 10:906. doi: 10.3389/fpsyg.2019.00906

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., et al. (2002). Evaluation of a behavioral measure of risk taking: the balloon analogue risk task (BART). *J. Exp. Psychol. Appl.* 8, 75–84. doi: 10.1037//1076-898X.8.2.75

Levin, H. M. (2013). "The utility and need for incorporating noncognitive skills into large-scale educational assessments" in *The role of international large-scale assessments: Perspectives from technology, economy, and educational research*. eds. M. von Davier, E. Gonzalez, I. Kirsch, and K. Yamamoto, (Dordrecht, Netherlands: Springer) 67–86.

Lipnevich, A. A., MacCann, C., and Roberts, R. D. (2013). "Assessing noncognitive constructs in education: A review of traditional and innovative approaches" in *Oxford handbook of child psychological assessment.* Vol. 1, eds. D. H. Saklofske, C. R. Reynolds, and V. Schwean (New York, NY: Oxford University Press).

Makransky, G., Mortensen, E. L., and Glas, C. A. W. (2013). Improving personality facet scores with multidimensional computer adaptive testing: an illustration with the neo pi-r. *Assessment* 20, 3–13. doi: 10.1177/1073191112437756

Moreno-Ger, P., Burgos, D., Martínez-Ortiz, I., Sierra, J. L., and Fernández-Manjón, B. (2008). Educational game design for online education. *Comput. Hum. Behav.* 24, 2530–2540. doi: 10.1016/j.chb.2008.03.012

Ortner, T. M., and Schmitt, M. (2014). Advances and continuing challenges in objective personality testing. *Eur. J. Psychol. Assess.* 30, 163–168. doi: 10.1027/1015-5759/a000213

Ranger, J., and Ortner, T. M. (2011). Assessing personality traits through response latencies using item response theory. *Educ. Psychol. Meas.* 71, 389–406. doi: 10.1177/0013164410382895

Renno, M. P., and Shutts, K. (2015). Children's social category-based giving and its correlates: Expectations and preferences. *Dev. Psychol.* 51, 533–543. doi: 10.1037/a0038819

Saxler, P. K. (2016). The marshmallow test: delay of gratification and independent rule compliance. PhD thesis. Harvard University. Retrieved from: http://nrs.harvard.edu/urn-3:HUL.InstRepos:27112705 (Accessed September 16, 2019).

Segall, D. O. (2005). Computerized adaptive testing. *Encycl. Soc. Meas.* 1, 429–438. doi: 10.1016/b0-12-369398-5/00444-8

Shute, V. J. (2011). "Stealth assessment in computer-based games to support learning" in *Computer games and instruction*. eds. S. Tobias and J. D. Fletcher (Charlotte, NC: Information Age Publishers), 503–524.

Silk, J. S., Steinberg, L., and Morris, A. S. (2003). Adolescents' emotion regulation in daily life: links to depressive symptoms and problem behavior. *Child Dev.* 74, 1869–1880. doi: 10.1046/j.1467-8624.2003.00643.x

Simmering, V. R., Ou, L., and Bolsinova, M. (2019). "A cross-disciplinary look at non-cognitive assessments" in *Quantitative psychology.* Vol. 265, eds. M. Wiberg, S. Culpepper, R. Janssen, J. González, and D. Molenaar (Cham, Switzerland: Springer), 157–167.

Smithers, L. G., Sawyer, A. C. P., Chittleborough, C. R., Davies, N. M., Davey Smith, G., and Lynch, J. W. (2018). A systematic review and meta-analysis of effects of early life non-cognitive skills on academic, psychosocial, cognitive and health outcomes. *Nat. Hum. Behav.* 2, 867–880. doi: 10.1038/s41562-018-0461-x

Springer, A., Garcia-Gathright, J., and Cramer, H. (2018). "Assessing and addressing algorithmic bias-but before we get there" in *AAAI Spring Symposium*. Palo Alto, CA: Association for the Advancement of Artificial Intelligence.

Stochl, J., Böhnke, J. R., Pickett, K. E., and Croudace, T. J. (2016). An evaluation of computerized adaptive testing for general psychological distress: combining GHQ-12 and Affectometer-2 in an item bank for public mental health research. *BMC Med. Res. Methodol.* 16:58. doi: 10.1186/s12874-016-0158-7

Stone, A. A., and Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavorial medicine. *Ann. Behav. Med.* 16, 199–202. doi: 10.1093/abm/16.3.199

Tan, P. Z., Forbes, E. E., Dahl, R. E., Ryan, N. D., Siegle, G. J., Ladouceur, C. D., et al. (2012). Emotional reactivity and regulation in anxious and nonanxious youth: a cell-phone ecological momentary assessment study. *J. Child Psychol. Psychiatry* 53, 197–206. doi: 10.1111/j.1469-7610.2011.02469.x

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., et al. (2014). "StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones" in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Seattle, WA: ACM), 3–14.

Wang, R., Harari, G., Hao, P., Zhou, X., and Campbell, A. T. (2015). "SmartGPA: how smartphones can assess and predict academic performance of college students" in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan: ACM), 295–306.

Woodzicka, J. A., and LaFrance, M. (2001). Real versus imagined gender harassment. *J. Soc. Issues* 57, 15–30. doi: 10.1111/0022-4537.00199

Conflict of Interest: VS, LO, and MB were employed by the non-profit company ACT, Inc.

frontiers in Psychology

ORIGINAL RESEARCH
published: 22 October 2019
doi: 10.3389/fpsyg.2019.02358

Combining Text Mining of Long Constructed Responses and Item-Based Measures: A Hybrid Test Design to Screen for Posttraumatic Stress Disorder (PTSD)

Qiwei He[1]*, *Bernard P. Veldkamp*[2], *Cees A. W. Glas*[2] *and Stéphanie M. van den Berg*[2]

[1] Educational Testing Service, Princeton, NJ, United States, [2] Department of Research Methodology, Measurement and Data Analysis, Faculty of Behavioural, Management and Social Sciences, University of Twente, Enschede, Netherlands

OPEN ACCESS

Edited by:
Frank Goldhammer,
German Institute for International Educational Research (LG), Germany

Reviewed by:
Alexander Robitzsch,
University of Kiel, Germany
Margot Mieskes,
Darmstadt University of Applied Sciences, Germany

****Correspondence:***
Qiwei He
qhe@ets.org

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *20 December 2018*
Accepted: *03 October 2019*
Published: *22 October 2019*

Citation:
He Q, Veldkamp BP, Glas CAW and van den Berg SM (2019) Combining Text Mining of Long Constructed Responses and Item-Based Measures: A Hybrid Test Design to Screen for Posttraumatic Stress Disorder (PTSD). Front. Psychol. 10:2358. doi: 10.3389/fpsyg.2019.02358

This article introduces a new hybrid intake procedure developed for posttraumatic stress disorder (PTSD) screening, which combines an automated textual assessment of respondents' self-narratives and item-based measures that are administered consequently. Text mining technique and item response modeling were used to analyze long constructed response (i.e., self-narratives) and responses to standardized questionnaires (i.e., multiple choices), respectively. The whole procedure is combined in a Bayesian framework where the textual assessment functions as prior information for the estimation of the PTSD latent trait. The purpose of this study is twofold: first, to investigate whether the combination model of textual analysis and item-based scaling could enhance the classification accuracy of PTSD, and second, to examine whether the standard error of estimates could be reduced through the use of the narrative as a sort of routing test. With the sample at hand, the combination model resulted in a reduction in the misclassification rate, as well as a decrease of standard error of latent trait estimation. These findings highlight the benefits of combining textual assessment and item-based measures in a psychiatric screening process. We conclude that the hybrid test design is a promising approach to increase test efficiency and is expected to be applicable in a broader scope of educational and psychological measurement in the future.

Keywords: posttraumatic stress disorder, text mining, item response theory, Bayesian framework, self-narratives

INTRODUCTION

Epidemiological research on mental illnesses such as posttraumatic stress disorder (PTSD) requires efficient methods to identify cases in large population-based samples (Shrout and Yager, 1989) because the diagnosis of the disorder is difficult to make and can involve expensive testing. A two-phase design can help on both accounts. The first phase involves a screening measure, meaning a more detailed diagnostic procedure needs to be administered solely to a selected subsample (Diamond and Lilienfeld, 1962; Shrout et al., 1986).

Item-based self-report instruments are often considered efficient for PTSD screening, as they usually require short administration time and do not require the presence of a clinician (Wohlfarth et al., 2003). Questionnaires such as the Trauma Assessment of Adults (Gray et al., 2009), the Brief Trauma Questionnaire (Schnurr et al., 2002), the Life Events Checklist (Gray et al., 2004), and the Trauma Life Events Questionnaire (Kubany et al., 2000) all have psychometric support for evaluating exposure to potentiality traumatic events. In addition to trauma exposure screeners, abbreviated PTSD symptom screeners are frequently used to determine the need for more in-depth clinical interviews (Lancaster et al., 2016). These include the Primary Care PTSD Screen (PC-PTSD; Prins and Ouimette, 2004), the Short Form of the PTSD Checklist-Civilian Version (Lang and Stein, 2005), the Trauma Screening Questionnaire (TSQ; Brewin et al., 2000), and the Short Post-Traumatic Stress Disorder Rating Interview (SPRINT; Connor and Davidson, 2001). These instruments ideally contain the minimal number of items necessary for accurate case identification, have simple decision rules to determine who passes and fails the screening, and are applicable to populations with varying prevalence of PTSD and experiencing different traumas (see more in reviews by Brewin, 2005; Lancaster et al., 2016).

As an alternative to such questionnaire-based screening, He et al. (2012) developed a computerized textual assessment system using text mining techniques, which was proved to be effective in analyzing open-ended writings regarding participants' trauma history and physical symptoms. The main idea was to analyze the respondents' textual input – the self-narratives describing traumatic experiences and impacts on their personal life to predict the risks of developing PTSD. In their study, the textual screening procedure resulted in a good agreement (82%) compared with a clinical structured interview in identifying the presence and absence of PTSD and yielded a higher sensitivity and positive prediction power than an itemized screening instrument.

With a growing body of research in learning patterns of language usage in psychiatric patients, textual input became recognized as an important additional source in the prediction of mental health (Pennebaker et al., 2003). For instance, Pennebaker (2001) found that linguistic markers, such as the use of negative-emotion words, cognition words, and insight words, predicted the future mental health of college students who wrote about traumatic events. Alvarez-Conrad et al. (2001) defined the presence of words relating to death and dying as an indicator of treatment-resistant PTSD. Consequently, the analysis of respondents' textual input and linguistic elements might provide crucial information for understanding cognitive mechanisms associated with trauma and hold valuable potential to screen for and predict PTSD symptoms and subtypes. Properly developed technologies such as text mining are expected to help individuals to self-test and public health organizations to screen for possible mental health conditions and prompt further evaluation when warranted, potentially preventing disorders from becoming chronic, debilitating, and difficult to treat (Todorov et al., 2018).

The focus of this study is to assess to what extent text mining techniques can be applied in the PTSD screening phase and to establish the extent to which they result in better estimates and better prediction of true diagnosis compared to the use of a questionnaire alone. Specifically, we propose a two-stage hybrid test design using a Bayesian approach to combine text mining and item response modeling in one systematic framework, where an automated score based on textual analysis serves as input for a prior distribution of a latent trait associated with PTSD that is measured by a number of questionnaire items using an item response theory (IRT) model (Rasch, 1960; Lord, 1980). Bayesian methods are especially useful for the estimation of a hierarchical structure (refer to Mislevy, 1986; Zwinderman, 1991), which allows extra prior information to be added into the measurement with the aim to increase prediction accuracy. Models developed in the Bayesian framework have been applied broadly in psychological and educational assessments. For instance, Matteucci and Veldkamp (2013) integrated students' background variables, such as scores obtained by the examinees from other tests, socioeconomic variables, and demographic variables as prior information to improve the accuracy of students' ability estimates (van den Berg et al., 2013) combined self-report and clinical interview data in a Bayesian approach to increase measurement precision in identifying schizotypal symptoms. However, the inclusion of textual assessments as prior information has been rarely described in the literature.

The purpose of this study is twofold: first, to investigate whether the combination model of textual analysis and item-based scaling can enhance the classification accuracy of PTSD, and second, to examine whether the standard error of estimates could be reduced through use of narrative as a kind of routing test. To examine the performance of our proposed method, we conducted a study to compare the estimates for a latent trait associated with PTSD with and without the use of a text mining score by means of three approaches: (1) an IRT-based test only, (2) textual analysis only, and (3) a combination of textual analysis and IRT-based itemized test including using the whole range of IRT-based items at one time and adding items adaptively starting from the one with the highest information, which is similar to the item selection procedure used in computerized adaptive testing (van der Linden and Glas, 2000).

MATERIALS AND METHODS

Sample and Instrument

Data used in the current study were collected from 105 trauma survivors via an online survey embedded in an open forum that is dedicated to people with mental health issues. Before administering items from the survey, all the participants were asked to report whether they had been diagnosed as PTSD or non-PTSD by psychiatrists via structured interviews with standardized instruments. Cases with missing diagnoses were discarded in the present study. Participants were also informed that the objective of the research was to develop a more flexible intake procedure for PTSD diagnosis and were requested to give responses to all the questions following the instructions.

The online survey consisted of two parts: self-narrative writing and administration of dichotomous questions regarding

PTSD symptoms. In the writing section, respondents were asked to write about their traumatic events and briefly describe the symptoms related to these experiences. Text length was recommended to be over 150 words, which was found as the average length of self-narratives input by PTSD patients in a previous study (He et al., 2012). In the item-based section, respondents were required to give compulsory answers to 21 items that were employed exactly the same in the National Comorbidity Study-Replication (NCS-R; Kessler et al., 2004) PTSD screening section. The NCS-R, conducted between February 2001 and April 2003 in the United States, is a nationally representative community household survey of the prevalence and correlates of mental disorders. These 21 dichotomous items (i.e., "yes" = 1, "no" = 0) one-to-one correspond to the PTSD symptoms that were defined in Diagnostic and Statistical Manual of Mental Disorders Fourth Version (DSM-IV; American Psychiatric Association, 2000). The first two columns in **Table 1** show the PTSD diagnostic criteria in the DSM-IV and their corresponding items that were used in the NCS-R as well as in this study.

Six of the 105 participants were excluded: Two reported they had never experienced traumatic events that were listed in the NCS-R, and four gave responses only to the item section but missed the writing section. This resulted in a total of 99 participants for the final set, among whom 34 were diagnosed as PTSD and 65 as non-PTSD. The sample had an age range between 19 and 63, with a mean of 30.06 (*SD* = 11.30). The majority of participants were female (78.4%). Over 90% participants had a higher educational background (i.e., college/university or above). 52.6% participants were reported as single, 40.2% were married, and 6.2% were divorced.

Procedure

To examine the performance of the hybrid test design, we estimated individuals' PTSD latent traits via three approaches: (1) an IRT-based test only, (2) text classification of self-narratives, and (3) combining textual analysis and IRT in a Bayesian framework. There were two analytic paths involved in the third approach: In one path, we combined the textual analysis with the whole set of 21 IRT-based items at a single time. In the other, we combined the textual analysis and the IRT latent scale in an

TABLE 1 | Item Parameters of 21 Questions Related to PTSD in NCS-R (calibrated with *n* = 880).

Item	Question in NCS-R	α	SE (α)	β	SE (β)	*r*
A2	Did you feel terrified or very frightened, helpless, shocked or horrified, numb at the time?	1.19	0.41	−4.45	0.48	0.19
B1	Did you ever have repeated unwanted memories of the event, that is, you kept remembering it even when you didn't want to?	1.82	0.20	−1.74	0.15	0.58
B2	Did you ever have repeated unpleasant dreams about the event?	1.24	0.14	−0.49	0.10	0.51
B3	Did you have flashbacks, that is, suddenly act or feel as if the event were happening over again?	1.41	0.15	−0.22	0.10	0.54
B4	Did you get very upset when you were reminded of the event?	1.64	0.18	−1.18	0.12	0.56
B5	When you were reminded of the event, did you ever have physical reactions like sweating, your heart racing, or feeling shaky?	1.68	0.17	−0.34	0.11	0.58
C1	After the event, did you try not to think about it?	0.95	0.12	−1.31	0.11	0.42
C2	After the event, did you purposely stay away from places, people or activities that reminded you of it?	1.34	0.14	−0.45	0.10	0.52
C3	After the event, were you ever unable to remember some important parts of what happened?	0.83	0.10	0.58	0.08	0.39
C4	After the event, did you lose interest in doing things you used to enjoy?	1.53	0.15	−0.39	0.10	0.53
C5	After the event, did you feel emotionally distant or cut-off from other people?	1.55	0.16	−0.88	0.11	0.53
C6	After the event, did you have trouble feeling normal feelings like love, happiness, or warmth toward other people?	1.86	0.18	−0.55	0.12	0.58
C7	After the event, did you feel you had no reason to plan for the future because you thought it would be cut short?	1.45	0.15	1.22	0.12	0.47
D1	During the time this event affected you most, did you have trouble falling or staying asleep?	1.14	0.18	−1.53	0.12	0.39
D2	During the time this event affected you most, were you more irritable or short-tempered than you usually are?	1.11	0.14	−0.16	0.09	0.46
D3	During the time this event affected you most, did you have more trouble concentrating or keeping your mind on what you were doing?	1.47	0.19	−1.10	0.11	0.48
D4	During the time this event affected you most, were you much more alert or watchful, even when there was no real need to be?	0.96	0.16	−0.85	0.10	0.39
D5	During the time this event affected you most, were you more jumpy or easily startled by ordinary noises?	1.28	0.17	−0.55	0.10	0.49
E1	Was any of these reactions continue to have at least 1 month?	0.78	0.30	−3.30	0.21	0.21
F1	Did these reactions cause distress to you?	1.55	0.26	−2.15	0.17	0.38
F2	Did these reactions disrupt or interfere with your normal, daily life?	1.02	0.16	−0.88	0.11	0.40

The item parameters were estimated from unidimensional 2PL model on a sample of 880 respondents in the NCS-R. SE indicates the standard error of item parameter estimation. r indicates validity coefficients that are calculated as the correlation of total score with each criterion item.

adaptive way, that is, we added the 21 items into the analysis one by one in descending order of item information available. We will illustrate each approach in detail in the following subsections. All analyses in the Bayesian framework were conducted using the software WinBUGS 1.4.3 (Lunn et al., 2000).

Approach 1: Using an IRT-Based Test Only

The IRT framework has been increasingly applied in psychiatric assessments in recent decades (e.g., van Groen et al., 2010; Weisscher et al., 2010; He et al., 2014b). In contrast to the classical sum score methods, IRT models (Rasch, 1960; Lord, 1980) provide improvement and flexibility by scaling the difficulty of items and the latent trait level of people on the same metric. Namely, the severity of prescribed symptoms and the latent degree of individuals' mental illness are set on a common scale, and thus can be meaningfully compared.

In the first approach, we focused on applying an IRT model on responses to the 21 PTSD diagnostic items in the NCS-R without adding any prior information. We employed a set of fixed item parameters that were previously calibrated using a larger sample size of 880 respondents collected in the NCS-R (He et al., 2014b). Note, however, that these 880 respondents gave responses to the questionnaire only, without any input by way of self-narratives. Given the objective of this study – examining the role of textual information in latent trait estimation to screen for PTSD, we had to collect a new sample of 99 respondents in this study who gave responses to both textual self-narratives and an itemized questionnaire, thus making it possible to combine both structured and unstructured data analysis in one framework.

In He et al. (2014b), given that symptom domains defined by the DSM-IV were used to index a general level of PTSD severity, we first considered a unidimensional two-parameter logistic (2PL) model underlying responses to the 21 symptoms (i.e., all 21 items on a single dimension). Next, given that the major 17 symptoms (in criteria domains B, C, and D) are placed *a priori* into three separate criterion domains, we also considered a three-dimensional IRT model where each domain was associated with a separate dimension. In addition, a special version of the 2PL model – the Rasch model or one-parameter logistic (1PL) model (Rasch, 1960) where the item discrimination parameter is simply fixed as one – was also considered, since such a model is often used in clinical applications as well (e.g., Wong et al., 2007; Elhai et al., 2011).

In the unidimensional 2PL model, that is, the probability of a score in category "yes" ($X_{ni} = 1$) of item i is given by the item response function

$$P(X_{ni} = 1|\theta_n) = \frac{\exp\left[\alpha_i(\theta_n - \beta_i)\right]}{1 + \exp\left[\alpha_i(\theta_n - \beta_i)\right]}, \tag{1}$$

where θ_n is the latent PTSD level of person n, β_i is an item difficulty parameter representing the severity level of each diagnostic symptom, and α_i is an item discrimination parameter indicating the extent to which the item response is related to the latent θ-scale. Note that in the Rasch model, the discrimination parameter α_i is fixed as 1. In the multidimensional version of the 2PL model, the probability of a positive response depends on M latent variables, say $\theta_{n1}, \ldots, \theta_{nm}, \ldots, \theta_{nM}$. In the multidimensional case, in eq. 1, the product $\alpha_i\theta_n$ is replaced by $\sum\limits_m \alpha_{im}\theta_{\ nm}$.

The dimensionality and model fit were examined using two steps: a likelihood ratio-statistic and an item-oriented Lagrange multiplier (LM) test. First, the likelihood-ratio test of the 2PL model against the Rasch model yielded a value of the test statistic $\chi^2 = 78.53, df = 16, p < 0.001,$ while the multidimensional model against the unidimensional 2PL model yielded a value of $\chi^2 = 37.41, df = 3, p < 0.001$. It was concluded that the multidimensional model fit the data best, and the 2PL fit the data significantly better than the Rasch model. However, although using a more complex model generally results in better model fit, using a more parsimonious model might still lead to adequate data description.

To investigate this, a second approach was used. Under each model, item fit was evaluated using an LM item fit statistic (Glas, 1998, 1999). These statistics can be used to evaluate the fit of the expected item response function given by Formula (1) to the observed item responses. Item fit was tested with a significance level of 0.01. For the Rasch model, the test was significant for six items, while no tests were significant for either the 2PL model or the multidimensional model. Further, the LM test statistic is accompanied by an effect size that measures the difference in observed and expected average item responses. For the 2PL model and the multidimensional model, these differences had the same magnitude. Hence, although a multidimensional IRT model fit the data better than 2PL in terms of the likelihood ratio test, it was not clearly superior in item fit. Therefore, the simpler unidimensional 2PL model was preferred over the more complicated multidimensional one. Consequently, the item calibration in the NCS-R was undertaken with the unidimensional 2PL model by marginal maximum likelihood (Bock and Aitkin, 1981) on a sample of 880 respondents in He et al. (2014b).

Further, we calculated validity coefficients r to examine how strong each criterion weighed on the general trait of PTSD and check whether these external criteria could match the discrimination parameters derived from the 2PL that indicates the extent to which the item response was related to the latent θ-scale. The validity coefficient is a statistical index used to report evidence of validity for intended interpretations of test scores and defined as the magnitude of the correlation between test scores and a criterion variable. We calculated the validity coefficients as the correlations between the NCS-R test results and each criterion variable and reported the results in the last column in **Table 1**. The larger the validity coefficient, the more confidence we can have in predictions made from the PTSD test scores. As shown in **Table 1**, the discrimination parameters in the third column showed a high agreement with the validity coefficients in the last column: for instance, the highest discrimination parameter located in criterion C6, where the top validity coefficient 0.58 was also found in this item. Similar findings were also applied to the lowest values of these two variables such as in criterion E1 and C3. The evidence demonstrated that the item weighting from the 2PL could provide similar conclusions based on external criteria (i.e., validity coefficients) to get consistent results in identifying strong (weak) factors in the test.

To maintain consistency with the previous study (He et al., 2014b), we fixed the calibrated item parameters in the current study. The fixed parameters and their standard errors were reported in the third column to the sixth column in **Table 1**. As shown here, the discrimination parameters varied in the interval [0.78, 1.86], with a mean value around 1.32. The difficulty parameters were included in the range [−4.45, 1.22], with a mean of −0.99. The respondents' latent traits were estimated by expected *a posteriori* (EAP) assuming a normal distribution.

Approach 2: Text Classification of Self-Narratives

Text classification is a special approach in the field of text mining, aiming to assign textual objects from a universe to two or more classes (Manning and Schütze, 1999). Supervised text classification generally involves two phases: a training phase and a testing phase. During the training phase, the most discriminative keywords to determine the presence or absence of PTSD are extracted and the relationship between the keywords and class labels is learned. The testing phase involves checking how well the trained classification model performs on a new dataset. In the testing procedure, each new input is scanned for the keywords that were extracted from training, and the most likely label for each new self-narrative is predicted. He et al. (2012) developed a supervised text classification model for PTSD screening. In this study, 300 self-narratives, consisting of 150 written by PTSD respondents and 150 written by non-PTSD respondents, were used to develop a screening system. In a follow-up study (He et al., 2017), four machine learning algorithms – including Decision Tree (DT), Naïve Bayes (NB), Support Vector Machine (SVM), and a self-developed alternative, the product score model (PSM) – were employed in conjunction with five data representations – unigrams, bigrams, trigrams, a combination of uni- and bigrams, and a mixture of n-grams. Unigram is the simplest and most commonly used data representation model where each word in a document collection acts as a distinct feature. N-gram considers the interaction effect of two, three, or more consecutive words (Manning and Schütze, 1999).

In He et al. (2017), it was found that narrative classification accuracy was maximized with the PSM in conjunction with unigrams. Although the addition of n-grams (i.e., bigrams and trigrams) has not significantly enhanced overall classification accuracy, it did help balance the performance metrics of text classification and improve the reliability of prediction. Furthermore, slight prevalence effects were found in the overall accuracy of all four machine learning algorithms; however, a substantial increase of positive prediction value (PPV) was noticed with the increase of prevalence of PTSD. When the prevalence of PTSD was low, the SVM and PSM had good sensitivity and high negative predictive power. This suggested that these two models could perform well in excluding the individuals identified as non-PTSD from the follow-up tests. Further, in a comparison with the mean performance of traditional screening measures reviewed by Brewin (2005), the SVM and PSM were shown to be more sensitive in detecting PTSD than the traditional screening measures, but their ability in detecting non-PTSD was a bit lower than the benchmark in clinical practice.

Because the PSM in conjunction with unigrams resulted in the highest agreement with the psychiatrists' diagnoses in clinical practice in the previous study (He et al., 2017), we applied this approach in the present study. We used the top 1,000 unigrams that were identified as the most robust classifiers to distinguish PTSD from the non-PTSD in He et al. (2012, 2017). Among the 1,000 unigrams, in descending order of word frequency, the 10 unique words most used by the PTSD patients were "rape," "flashback," "fire," "involve," "avoid," "incident," "date," "tower," "men," and "fault." The words "test," "hardly," "tumor," "tight," "excite," "evil," "pleasure," "vision," "frantic," and "funny" were found to be the top 10 in the non-PTSD corpus (He et al., 2012). Analogous to the results obtained by Orsillo et al. (2004) in the research regarding emotion expressions of PTSD patients, the words favored by PTSD patients had relatively stronger negative semantic tendency no matter the lexical form: adjective, noun, or verb (He et al., 2012).

A preprocessing routine was implemented to standardize the n-grams for textual analysis, which was consistent with the previous studies (He et al., 2012, 2017). This involved screening digital numbers, deducting non-informative "stop words"[1] (e.g., "*I*," "*to*"), common punctuation marks (e.g., "," ":") and frequently used abbreviations (e.g., "*isn't*," "I'm"), and "stemming" the rest of the words, using the Porter algorithm (Porter, 1980), to remove common morphological endings. For example, the terms "*nightmares,*" "*nightmaring,*" and "*nightmare*" were normalized in an identical stem "*nightmar*"[2] by removing the suffixes and linguistic rule-based indicators (for more preprocessing rules refer to Manning and Schütze, 1999; He et al., 2012, 2017).

The PSM is an alternative machine learning algorithm to address the smoothing issue of NB using a form of Laplace's law (Laplace, 1995). This model was validated in previous studies (He et al., 2012, 2017). Holding the similar independence assumption as the NB model, the PSM features assigning two weights for each keyword (in binary classification) to indicate how popular the keywords are in the corpora of self-narratives written by either PTSD patients (corpus[3] C_1) or non-PTSD patients (corpus C_2). The name *product score* comes from a product operation to compute scores for each class, that is, S_1 and S_2, for each input text based on the term weights. To be consistent with the previous studies, we used the smoothing constant $a = 0.5$, which was added to the word frequency to account for words that did not occur in the training set but might occur in new texts (for more smoothing rules refer to Manning and Schütze, 1999; Jurafsky and Martin, 2009). The equation is,

$$\begin{cases} S_1 = P(C_1) \cdot \prod_{w=1}^{k} \left[(u_w + a)/len(C_1) \right] \\ S_2 = P(C_2) \cdot \prod_{w=1}^{k} \left[(v_w + a)/len(C_2) \right], \end{cases} \tag{2}$$

[1] The current study used the standard "English Stop Word List" (127 words) in Python Natural Language Toolkit (NLTK, Perkins, 2010) to deduct the non-informative words.

[2] The stemming algorithm is used to normalize lexical forms of words, which may generate stems without an authentic word meaning, such as "*nightmar.*"

[3] A body of texts is usually called a text corpus.

where u_w and v_w are the number of occurrences of keyword w in both corpora C_1 (i.e., PTSD corpus) and C_2 (i.e., non-PTSD corpus), respectively. $len(C)$ is the corpus length, namely, the sum of the word occurrences in each corpus. $P(C)$ is the prior probability of a certain class in the whole corpus collection. The classification rule is defined as:

$$\text{choose} \begin{cases} C = 1 & \text{if } \log(S_1/S_2) > b \\ C = 2 & \text{else} \end{cases}, \tag{3}$$

where b is a constant set as zero in this study. The reason was that in the previous study (He et al., 2012) it was found during the PTSD textual screening procedure that the largest number of positive cases could be captured without unduly sacrificing specificity when the threshold was set at zero. The value of $\log(S_1/S_2)$ was defined as the text score for each self-narrative (see also He and Veldkamp, 2012; He et al., 2012). For an easy comparison with the IRT scales, we standardized the text scores as $Z \sim N(0, 1)$[4].

Approach 3: Combining Textual Analysis and IRT in a Bayesian Framework

Textual analysis and item response modeling were combined in a Bayesian framework, where the text score of each self-narrative obtained in approach 2 was used as prior information. The posterior distribution of the latent PTSD level is proportional to the product of the prior and the likelihood, that is,

$$P(\theta|x, y) \propto p(x|\theta, \alpha, \beta)g(\theta|y), \tag{4}$$

where x is the vector of responses to the questionnaire, y is the text score for each individual, $g(\theta|y)$is the prior given the covariate of textual assessments, α and β are the fixed discrimination and difficulty parameters of items, $p(x|\theta, \alpha, \beta)$ is the likelihood function of the IRT model. The relation between the PTSD latent trait θ of individual n and the text score y_n is given by the linear regression

$$\theta_n = b_0 + b_1 y_n + \varepsilon_n, \tag{5}$$

where b_0 and b_1 are the regression coefficients. The error terms are assumed to be independent and normally distributed as $\varepsilon_n \sim N(0, \sigma^2)$ with $n = 1, ..., N$ individuals. The assumption of a linear regression model is translated into a normal conditional distribution of θ_n given the text covariate as

$$\theta_n | y_n \sim N(b_0 + b_1 y_n,\ \sigma^2) \tag{6}$$

Formula (6) represents an informative prior distribution of the PTSD latent trait. For each individual, the estimation of latent trait was performed by using 5,000 Markov chain Monte Carlo (MCMC) iterations with the burn-in of length of 1,000.

To determine whether the introduction of the prior distribution was effective, we compared the posterior distribution of θ_n in the combination model with the estimation from the IRT-based test only. Because the item parameters in the IRT model were fixed, the θ-estimates resulting from both of the IRT-based test and the combination model (use textual information as a prior) were on a common scale and thus could be compared.

Two investigations were conducted to analyze the efficiency of the combination model. The first was to combine the textual assessments with the full range of 21 items of the NCS-R questionnaire. The main purpose was to explore whether adding the text prior would significantly impact the accuracy of PTSD detection. The second investigation pursued the question of whether adding textual assessments to the questionnaire could result in a reduction of the number of items administered without sacrificing precision of the θ-estimates. Those items that provide peak information around the cutoff threshold are ideal for a shorter version of a mastery test (Thomas, 2011). Since the target of screening is to make classification decisions, a natural choice would be to maximize information at the chosen diagnostic cutoff (for more about item information refer to Lord, 1980). In the current study, we employed the same cutoff point at $\theta = -0.15$ that was derived from He et al. (2014b) to distinguish PTSD and non-PTSD using a larger sample size of 880 respondents collected in the NCS-R. As mentioned above, this study shared the same questionnaire scale and item parameters as He et al. (2014b). This ensured the value of the cutoff point was comparable in these two studies. Further, the cutoff point derived based on a larger sample size was shown to be more reliable than a smaller sample size, so we kept the cutoff value consistent.

In He et al. (2014b), three approaches were used to set the standard (i.e., obtain a cutoff point on the latent scale) to distinguish PTSD and non-PTSD. The first approach entailed finding the midpoint between the medians of the two distributions (Cizek and Bunch, 2007). The second was the contrasting-groups method (Brandon, 2002), which uses logistic regression to determine the latent score point at which the probability of category membership is 50%. Setting the respondent status as a dichotomous variable coded 0 = non-PTSD and 1 = PTSD, we entered the latent scores of all the respondents into a general logistic regression equation; that is, $y^* = a + bx$, where y^* is the predicted value of the outcome variable (respondent status) for a respondent and x is the respondent's observed score. Given $y^* = 0.5$, the classification cutoff point for PTSD and non-PTSD groups could be obtained simply. The third approach used the Bayesian discrimination function, which minimizes expected risk. Using the zero-one loss function, the decision boundary becomes $g_i(x) = P(C_i|x) = \frac{P(C_i)p(x|C_i)}{p(x)}$, where $P(C_i)$ is the prior probability (i.e., the prevalence of PTSD or non-PTSD in the total sample); $p(x|C_i)$ represents the class likelihood (we assumed the latent trait scores have a normal distribution); and $p(x)$ indicates the marginal probability of observation x. Given the assumption of normal distribution in both PTSD and non-PTSD groups, we could derive the cutoff point. Finally, we calculated the average of these three cutoff points based on the 21 items in the NCS-R and got -0.15 as the cutoff point on the latent scale.

[4] In He et al. (2014b), the IRT parameters were calculated by the marginal maximum likelihood method with the assumption that ability was in a standard normal distribution. The original ability scale was therefore also in a standard normal distribution. In other words, after fixing the IRT parameters, the resulting ability scores are on a standard normally distributed scale. Therefore we can normalize the text score on the same scale.

Consequently, in the present study, we calculated the item information for all the 21 items at this derived cutoff point and ranked the items in a descending order, namely, starting from the item with the highest information to the least information (see **Figure 1**). The items were ordered as following: C6, B5, C4, B3, C5, C2, D5, B2, B4, D3, D2, F2, C7, D4, D1, C1, B1, C3, F1, E1, A2. We started to examine the performance of a combination of the text prior and the most informative item – text prior with item C6 (i.e., "did you have trouble feeling normal feelings like love, happiness, or warmth toward other people?") versus using item C6 alone. The second informative item (B5) was then added in for the comparison of the next pattern. The procedure continued until all the 21 items were included. Both test information and standard error of θ – estimates were calculated for each pattern (i.e., with and without text prior) with an increasing number of informative items. Since textual assessment was suggested as a sort of complementary information to predict people's physical and mental health (e.g., Gottschalk and Gleser, 1969; Rosenberg and Tucker, 1979; Smyth, 1998; Franklin and Thompson, 2005), the test information was expected to increase, and the standard errors were expected to decrease when text priors were added.

The performance of the three approaches was compared on five metrics: accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). The diagnoses made in the structured interviews by psychiatrists were used as the true standard in the comparison. Accuracy, the main metric used in classification, is the percentage of correctly defined individuals. Sensitivity and specificity are the proportion of actual positives and actual negatives that are correctly identified, respectively. These two indicators do not depend on the prevalence in the sample (i.e., proportion of "PTSD" and "non-PTSD" of the total), and hence are indicative of real-world performance. The predictive values, PPV and NPV, are estimators of the confidence in predicting correct classification, that is, the higher predictive values are, the more reliable the prediction is.

RESULTS

For the sample of 99 participants, the latent trait estimation via approach 1 resulted in a normal distribution of latent trait levels θ_n, with a mean value of −0.39 and variance of 2.31. The standardized text scores obtained from approach 2 resulted in a range [−2.92, 4.22]. In approach 3, the latent linear regression model given by Formula (4) and (5) was estimated using the item responses and the textual covariates. The intercept and slope coefficients were obtained as −0.41 and 1.44, respectively. The error term in the prior information (textual covariates) had a normal distribution with a mean value of zero and variance of 3.57. Hence, the informative prior distribution of the PTSD latent trait was defined as $\theta_n | y_n \sim N(-0.41 + 1.44y_n, 3.57)$.

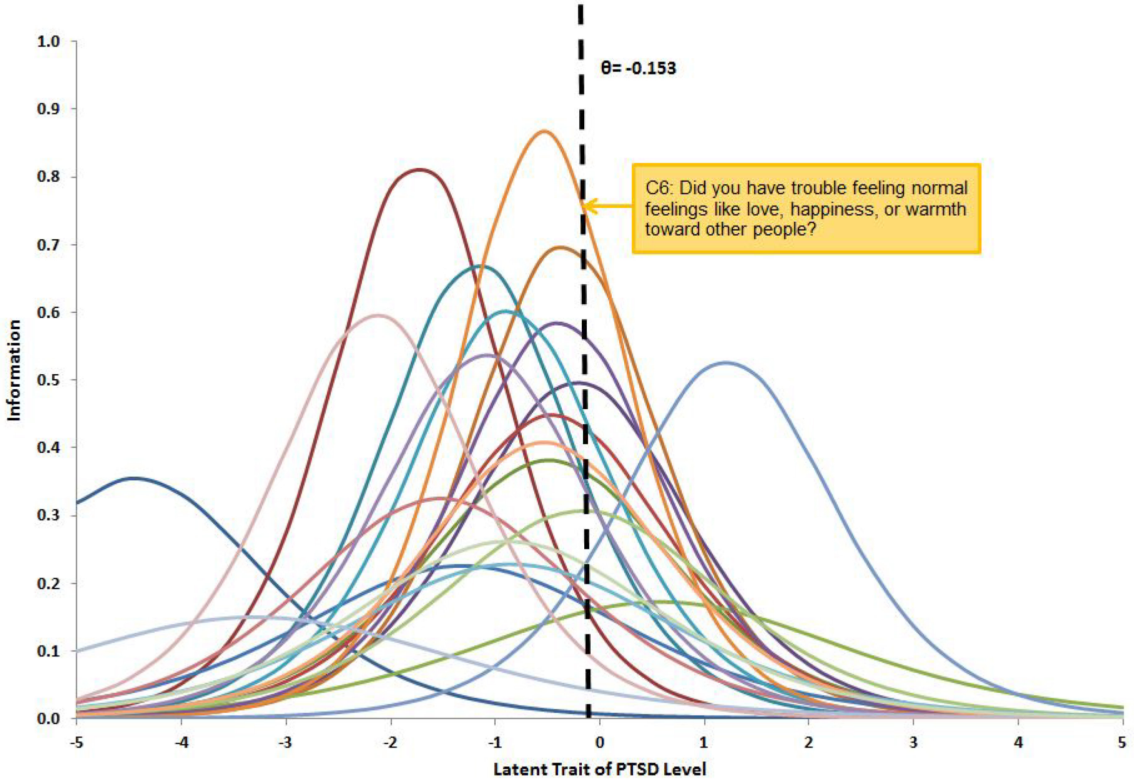


FIGURE 1 | Item information for 21 items in NCS-R questionnaire corresponding to DSM-IV PTSD diagnosis criteria. The cutoff point was estimated at −0.15 on latent scale to distinguish PTSD and non-PTSD. Item C6 is the most informative item, having the highest intersection value with the cutoff line.

The correlations among the estimations from the three approaches are presented in **Table 2**. It was noted that the correlation between the EAP of θ-estimates via approach 1 and the text scores estimated via approach 2 was 0.56, suggesting that there was a positive and moderate relation between the self-narrative writing and the responses to the itemized questionnaire in the structured interview. This result reiterated the findings in the earlier studies that the words and expressions were capable of predicting one's mental health status.

Table 3 shows the performance metrics of the three approaches. As we expected, the diagnostic accuracy rate was fairly high – 0.94 – when using the 21-item questionnaire by the IRT alone, and was improved to 0.97 with an addition of textual assessment. It suggested that 6 out of 99 respondents were misclassified using the IRT scale alone, while the misclassification rate decrease to 3 out of 99 respondents when adding the textual analysis as prior information. Using a 95% confidence interval, the paired sample *t*-test showed that the mean of latent trait estimation ($t = 3.86$, $df = 98$, $p < 0.01$) and standard deviation of latent trait distribution both significantly differ with and without text prior ($t = 3.70$, $df = 98$, $p < 0.01$). That is, the extra information gained from the textual analysis helped the latent trait locate closer to their true value, which helped decrease the misclassification rate by 50%. Given concerns on only using the keywords as predictors to make the classification, the accuracy rate (0.84) produced by the textual assessment was satisfactorily high, although it was a bit lower than the other two approaches. The sensitivity and NPV were perfect for all three approaches, implying that both the IRT and the textual assessments were sensitive for identifying PTSD patients. With the introduction of textual assessment, the specificity and PPV rose to 0.95 and 0.92, respectively. It suggested that the textual assessment played an effective role in detecting non-PTSD and strengthened the power in identifying PTSD in the population.

We further examined the relationship between the standard error of the estimate of θ and the number of items with the presence or absence of text prior. We added in items into the analysis one by one following an adaptive way with a descending order of the item information, which was derived at the cutoff point introduced in the **Figure 1**. As shown in **Figure 2**, the horizontal axis indicates the number of items in the IRT model and the vertical axis indicates the average standard error of the latent trait estimation. The curve of standard error without using the text prior (i.e., the dash line), that is, using the IRT model alone via approach 1, starts around 1.6 and drops gradually to 0.68 when all the 21 items are included. The curve of standard error using a text prior (i.e., the solid line) follows a similar pattern but stays on a lower level than the dash curve. It starts around 1.4 (when the first item with the highest information was included) and ends around 0.65 (when all the 21 items were included). Using a 95% confidence interval, the paired sample *t*-test showed that the standard error of estimation with text prior was significantly lower than that without text prior ($t = 3.86$, $df = 98$, $p < 0.01$) when including the whole range of 21 items. With the increasing number of items, the differences between these two curves decreased from 0.20 to 0.03. It suggested that the textual assessment did have an impact on the latent trait estimation, and the effect was more apparent when using fewer items. The red dotted line highlights the standard error when using 21 items without the text prior. It crosses the solid curve at 17 items, implying that with the introduction of the text prior, 17 items would be good enough to make the estimation as precisely as using the whole range of 21 items. That is, by using the text priors, the questionnaire length can be shortened by 4 items without sacrificing precision.

TABLE 2 | Correlations among estimates from three approaches: IRT, TX, and a combination of TX and IRT (21-item).

	IRT	TX	TX and IRT (21-item)
IRT	1.00		
TX	0.56	1.00	
TX and IRT (21-item)	0.99	0.62	1.00

TX indicates the textual assessments. Correlation is significant at the 0.01 level (2-tailed).

TABLE 3 | Performance metrics compared among IRT, TX, and a combination of TX and IRT (21-item).

	Accuracy	Sensitivity	Specificity	PPV	NPV
IRT	0.94	1.00	0.92	0.87	1.00
TX	0.84	1.00	0.77	0.69	1.00
TX and IRT (21-item)	0.97	1.00	0.95	0.92	1.00

TX indicates textual assessment. PPV and NPV represent the positive predictive value and negative predictive value, respectively.

DISCUSSION

In this study, a new intake procedure for PTSD screening was developed that combined an automated textual assessment of patients' self-narratives and an itemized questionnaire. To determine whether the introduction of text information is effective, we identified PTSD cases via three approaches: (1) we estimated PTSD latent trait by using IRT on a standardized questionnaire, (2) classified patients' self-narratives into PTSD and non-PTSD groups by using a text mining technique, and (3) estimated the posterior distribution of PTSD latent trait by combining textual assessments and IRT in a Bayesian framework by both a linear and adaptive method. With the sample at hand, the results showed that the combination model enhanced the accuracy of PTSD detection from 0.94 to 0.97, reduced the standard error of latent trait estimation, and could shorten the questionnaire length by four items without sacrificing accuracy.

In the current study, the diagnostic accuracy was already high (0.94) when using the itemized questionnaire alone (approach 1). However, a structured interview that generally employs questionnaires is time consuming in daily practice. The computerized textual assessment proposed in this study is relatively easy to conduct via the internet. The highly satisfactory detection accuracy rate (0.84) is promising for real application. Note that the threshold in textual analysis could be adjusted according to the requirements of the practioner, for instance,

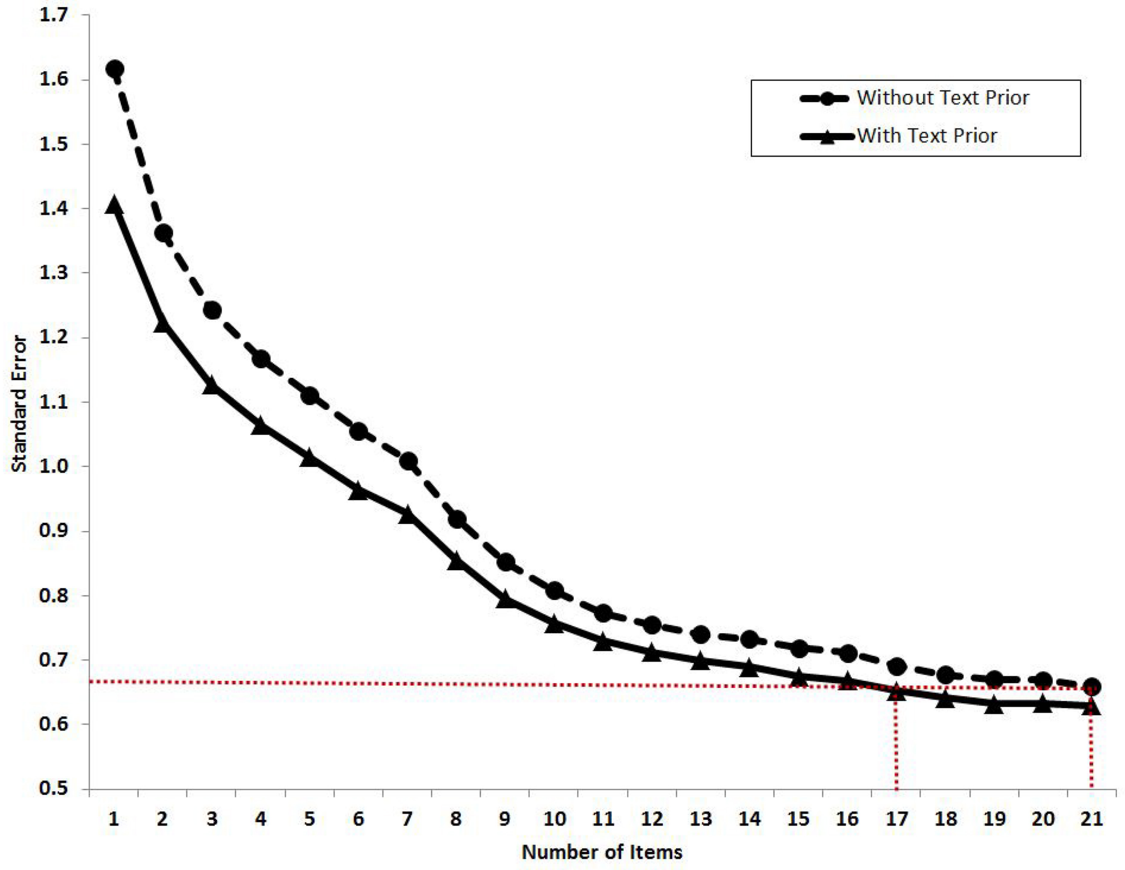


FIGURE 2 | The relationship between standard error of the estimate of θ and the increasing number of items with or without using text priors. The red dotted line indicates the standard error when using 21 items without text priors. It crosses the solid standard error curve at 17 items, meaning that by using the text priors, the test length can be shortened by four items. The order of items is ranked by a descending order of item information with the cutoff point that was derived in **Figure 1**.

using a relatively lower threshold to include the maximum number of PTSD potential patients for the second step in an itemized questionnaire, or increasing the threshold to a higher value in order to precisely detect PTSD patients by the textual assessment alone (He et al., 2012). Given concerns of the cost-effectiveness of the screening at an initial stage, it would be interesting to combine these two approaches in a two-phase framework to reduce clinical expense and improve the accuracy rate.

Further, according to the results in the previous study of He et al. (2012), the NPV of the textual assessments was satisfactorily high – 0.85 – when the text classification algorithm PSM was applied in conjunction with unigrams. It meant that the textual screening tool was helpful in excluding the non-PTSD respondents from the follow-up tests. For the 99 sample in the present study, taking the 85% confidence interval, 53 respondents could be excluded from the further tests.

It is also worthwhile to discuss the cost-effectiveness of the hybrid test design that combined the textual analysis and item-based test. The results showed that using textual information helped save follow-up items. However, weighing the benefits of the text prior, we would also take the amount of time it takes to write self-narratives into account. On the one hand, from respondents' perspective, writing self-narratives provides flexibility to express the individual's inner world and prevents being passively triggered by sensitive questions, even if the process might take longer than directly responding to the itemized questionnaire. On the other hand, from the practitioners' perspective, the procedure for item development is often time consuming and involves multiple steps (e.g., data collection, data cleaning, field trial, item parameter calibration, and examination of reliability and validity of a scale). Comparatively, textual analysis could substantially shorten scale-development time and simplify the procedure once the model is successfully trained and refined with different textual contexts.

In addition, structured textual analysis that usually involves tight structures from existing software, such as Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2001), is a good supplement to the text mining-based techniques. LIWC is a textual analysis software program that looks for words and counts them in categories relevant to psychology across multiple text files, for instance, essays, emails, blogs, novels, and so on. It has two central features – the processing component and dictionaries. During processing, the program goes through each file word by word. Each word in a given text file is compared with the dictionary file. A dictionary refers to the collection of words that define a particular category such as "family," "positive emotion," and "work." In a pilot study based on 50 self-narratives, half written by a PTSD group and half by a non-PTSD group, it was found that the PTSD respondents used significantly more emotional words and expressions related to family. These results are interesting enough to be addressed in another paper in the future.

Some limitations in the present study also merit discussion. First, the sample size was rather small at only 99 participants. Second, it was notable that female respondents represented the majority (approximately 78%) in the sample, which was consistent with the proportion of females in the target sample of PTSD[5] in the NCS-R. Further, evidence has shown that females are associated with a higher risk for PTSD (e.g., Lancaster et al., 2016). It would be interesting to examine whether the screening method (with text priors) plays an equal role in detecting PTSD in males and females, especially given concerns about the potential differences in their writing habits. Third, those in the sample had an unusually high level of education. This was probably caused largely by data collection being conducted on an internet platform. People with a higher educational background are possibly easier accessed via a web-based test than a less educated group (Naglieri et al., 2004). It would be interesting to make a comparative study in the future to investigate whether demographic variables (e.g., age, gender, and education) could make an impact on the textual assessment and hybrid model.

Last but not least, since the data used in this study was collected via an online platform, special caution needs to be taken as far as the potential risk of fake information. We had invited at least two psychiatrists to check each self-narrative entry to ensure the input was reasonable and authentic and could be used in this study. However, how to validate the internet data before entering data processing would be an important topic. For instance, He et al. (2014a) introduced an approach to detect potential fake information on social media (i.e., Facebook) data collection via statistical models on person and item fit.

Prevalence of a condition is an important indicator when reporting the performance metrics of a screening method. Whereas sensitivity and specificity are independent of the prevalence of the disorder in the population, positive and negative predictive power are sensitive to population prevalence (Brewin, 2005). In our previous study (He et al., 2014b), we reported the possible prevalence as ranging from 5 to 50% and noticed that there was little difference in the accuracy of screening for PTSD using the PSM model when the range of prevalence was so large. It was also noticed that when the prevalence of PTSD in the sample was increased, the PPV increased as well. It meant that the confidence of correctly identifying PTSD also increased. In the current study, we note that both specificity and PPV increased when we used the hybrid model.

In summary, the current study presented a new trial in developing a hybrid model to combine textual assessment of patients' self-narratives and itemized questionnaire in detecting mental illness. Its aim was to reduce the respondents' burden and clinicians' workload. Adding textual prior information, detection accuracy could be enhanced and test length could be shortened. The results demonstrated that the combination of a textual assessment and an IRT-based questionnaire is a promising approach to increase cost-effectiveness in PTSD diagnosis and is expected to be applicable in a broader scope of both (online) screening and psychiatric diagnosis as well as other psychological and educational assessments in the future. Further, with the rapid development of computer-based assessments, more data could be captured during the assessment process. The use of timing data as well as action sequences, keystrokes (e.g., type in and delete), and other process-related information hold promise for contributions to the advancement of screening methods in future research.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of Code of Ethics for Research in the Social and Behavioural Sciences Involving Human Participants used as the guidelines by the Faculty of Behavioural, Management and Social Sciences (BMS) Ethics Committee, University of Twente with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the BMS Ethics Committee, University of Twente.

AUTHOR CONTRIBUTIONS

QH contributed to the development of the methodological framework and the model estimation procedures, conduction of the data analysis, and the drafting and revision of the manuscript. BV contributed to providing suggestions on the methodological framework and the model estimation procedures, and the reviewing and revision of the manuscript. CG contributed to providing suggestions on the methodological framework, and the reviewing of the manuscript. SB contributed to providing suggestions on the model estimation procedures and conduction of the data analysis, and the reviewing of the manuscript.

FUNDING

This study was partially supported by the Stichting Achmea Slachtoffer en Samenleving, Netherlands.

ACKNOWLEDGMENTS

This manuscript was included in the first author's doctoral thesis "Text Mining and IRT for Psychiatric and Psychological Assessment" (He, 2013). The authors confirmed this is the only medium it has appeared in and is in line with the authors' university policy. The thesis can be accessed online at https://research.utwente.nl/en/publications/text-mining-and-irt-for-psychiatric-and-psychological-assessment. The authors would like to thank Larry Hanover for his help in reviewing this manuscript.

[5] Only people who had mental health problems or were screened as positively high potential into mental problems in the round 1 were included as a target sample of PTSD in the NCS-R.

REFERENCES

Alvarez-Conrad, J., Zoellner, L. A., and Foa, E. B. (2001). Linguistic predictors of trauma pathology and physical health. *Appl. Cogn. Psychol.* 15, 159–170.

American Psychiatric Association, (2000). *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*. Washington, DC: Author.

Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM-algorithm. *Psychometrika* 46, 443–459. doi: 10.1007/bf02293801

Brandon, P. R. (2002). Two versions of the contrasting-groups standard-setting method: a review. *Measur. Eval. Counsel. Dev.* 35, 167–181. doi: 10.1080/07481756.2002.12069061

Brewin, C. R. (2005). Systematic review of screening instruments for adults at risk of PTSD. *J. Trauma. Stress* 18, 53–62. doi: 10.1002/jts.20007

Brewin, C. R., Andrews, B., and Valentine, J. D. (2000). Meta-analysis of risk factors for posttraumatic stress disorder in trauma-exposed adults. *J. Consult. Clin. Psychol.* 68, 748–766. doi: 10.1037/0022-006x.68.5.748

Cizek, G. J., and Bunch, M. B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA: Sage.

Connor, K., and Davidson, J. (2001). Sprint: a brief global assessment of post-traumatic stress disorder. *Int. Clin. Psychopharmacol.* 16, 279–284. doi: 10.1097/00004850-200109000-00005

Diamond, E. L., and Lilienfeld, A. M. (1962). Effects of errors in classification and diagnosis in various types of epidemiological studies. *Am. J. Public Health* 52, 1137–1144. doi: 10.2105/ajph.52.7.1137

Elhai, J. D., de Francisco Carvalho, L., Miguel, F. K., Palmieri, P. A., Primi, R., and Frueh, B. C. (2011). Testing whether postraumatic stress disorder and major depressive disorder are similar or unique constructs. *J. Anxiety Disord.* 25, 404–410. doi: 10.1016/j.janxdis.2010.11.003

Franklin, C. L., and Thompson, K. E. (2005). Response style and posttraumatic stress disorder (PTSD): a review. *J. Trauma Dissociation* 6, 105–123. doi: 10.1300/j229v06n03_05

Glas, C. A. W. (1998). Detection of differential item functioning using lagrange multiplier tests. *Stat. Sin.* 8, 647–667.

Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika* 64, 273–294. doi: 10.1007/BF02294296

Gottschalk, L. A., and Gleser, G. C. (1969). *The Measurement of Psychological States Through the Content Analysis of Verbal Behavior*. Berkeley, CA: University of California Press.

Gray, M. J., Elhai, J. D., Owen, J. R., and Monroe, R. (2009). Psychometric properties of the trauma assessment for adults. *Depress. Anxiety* 26, 190–195. doi: 10.1002/da.20535

Gray, M. J., Litz, B. T., Hsu, J. L., and Lombardo, T. W. (2004). Psychometric properties of the life events checklist. *Assessment* 11, 330–341. doi: 10.1177/1073191104269954

He, Q. (2013). *Text Mining and IRT for Psychiatric and Psychological Assessment*. Enschede: Universiteit Twente, doi: 10.3990/1.9789036500562

He, Q., Glas, C. A. W., Kosinski, M., Stillwell, D. J., and Veldkamp, B. P. (2014a). Predicting self-monitoring skills using textual posts on Facebook. *Comput. Hum. Behav.* 33, 69–78. doi: 10.1016/j.chb.2013.12.026

He, Q., Glas, C. A. W., and Veldkamp, B. P. (2014b). Assessing the impact of differential symptom endorsement on posttraumatic stress disorder (PTSD) diagnosis. *Int. J. Methods Psychiatr. Res.* 23, 131–141. doi: 10.1002/mpr.1417

He, Q., and Veldkamp, B. P. (2012). "Classifying unstructured textual data using the product score model: an alternative text mining algorithm," in *Psychometrics in Practice at RCEC*, eds T. J. H. M. Eggen, and B. P. Veldkamp, (Enschede: RCEC), 47–62.

He, Q., Veldkamp, B. P., and de Vries, T. (2012). Screening for posttraumatic stress disorder using verbal features in self narratives: a text mining approach. *Psychiatry Res.* 198, 441–447. doi: 10.1016/j.psychres.2012.01.032

He, Q., Veldkamp, B. P., Glas, C. A. W., and de Vries, T. (2017). Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment* 24, 157–172. doi: 10.1177/1073191115602551

Jurafsky, D., and Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Pearson Prentice Hall.

Kessler, R. C., Berglund, P., Chiu, W. T., Demler, O., Heeringa, S., Hiripi, E., et al. (2004). The US national comorbidity survey replication (NCS-R) design and field procedures. *Int. J. Methods Psychiatr. Res.* 13, 69–92. doi: 10.1002/mpr.167

Kubany, E. S., Leisen, M. B., Kaplan, A. S., Watson, S. B., Haynes, S. N., Owens, J. A., et al. (2000). Development and preliminary validation of a brief broad-spectrum measure of trauma exposure: the traumatic life events questionnaire. *Psychol. Assess.* 12, 210–224. doi: 10.1037//1040-3590.12.2.210

Lancaster, C. L., Teeters, J. B., Gros, D. F., and Back, S. E. (2016). Posttraumatic stress disorder: overview of evidence-based assessment and treatment. *J. Clin. Med.* 5:105. doi: 10.3390/jcm5110105

Lang, A. J., and Stein, M. B. (2005). An abbreviated PTSD checklist for use as a screening instrument in primary care. *Behav. Res. Ther.* 43, 585–594. doi: 10.1016/j.brat.2004.04.005

Laplace, P. S. (1995). *Pierre-Simon Laplace Philosophical Essay on Probabilities*, Vol. 13. New York, NY: Springer.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, CA: Erlbaum.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS: a bayesian modeling framework: concepts, structure, and extensibility. *Stat. Comput.* 10, 325–337.

Manning, C. D., and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Matteucci, M., and Veldkamp, B. P. (2013). On the use of MCMC computerized adaptive testing with empirical prior information to improve efficiency. *Stat. Methods Appl.* 22, 243–267. doi: 10.1007/s10260-012-0216-1

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika* 51, 177–195. doi: 10.1007/bf02293979

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., et al. (2004). Psychological testing on the internet. *Am. Psychol.* 59, 150–162.

Orsillo, S. M., Batten, S. V., Plumb, J. C., Luterek, J. A., and Roessner, B. M. (2004). An experimental study of emotional responding in women with posttraumatic stress disorder related to interpersonal violence. *J. Trauma. Stress* 17, 241–248. doi: 10.1023/b:jots.0000029267.61240.94

Pennebaker, J. W. (2001). Dealing with a traumatic experience immediately after it occurs. *Adv. Mind Body Med.* 17, 160–162.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Linguistic Inquiry and Word Count*. Mahwah, NJ: Erlbaum.

Pennebaker, J. W., Mehl, M. R., and Niederhoffer, K. G. (2003). Psychological aspects of natural language. use: our words, our selves. *Annu. Rev. Psychol.* 54, 547–577. doi: 10.1146/annurev.psych.54.101601.145041

Perkins, J. (2010). *Python Text Processing with NLTK 2.0 Cookbook*. Birmingham: Packt Publishing Ltd.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program Autom. Library Inform. Syst.* 14, 130–137. doi: 10.1108/eb046814

Prins, A., and Ouimette, P. (2004). The primary care PTSD screen (PC-PTSD): development and operating characteristics. *Primary Care Psychiatry* 9, 9–14. doi: 10.1016/j.jad.2019.05.021

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

Rosenberg, S. D., and Tucker, G. J. (1979). Verbal-behavior and schizophrenia—semantic dimension. *Arch. Gen. Psychiatry* 36, 1331–1337.

Schnurr, P. P., Spiro, A., Vielhauer, M. J., Findler, M. N., and Hamblen, J. L. (2002). Trauma in the lives of older men: findings from the normative aging study. *J. Clin. Geropsychol.* 8, 175–187.

Shrout, P. E., Skodol, A. E., and Dohrenwend, B. P. (1986). "A two-stage approach for case identification and diagnosis, first stage instruments," in *Mental Disorders in Community: Progress and Challenge*, eds J. E. Barrett, and R. M. Rose, (New York, NY: Guilford), 286–303.

Shrout, P. E., and Yager, T. J. (1989). Reliability and validity of screening scales: effects of reducing scale length. *J. Clin. Epidemiol.* 42, 69–78. doi: 10.1016/0895-4356(89)90027-9

Smyth, J. M. (1998). Written emotional expression: effect sizes, outcome types and moderating variables. *J. Consult. Clin. Psychol.* 66, 174–184. doi: 10.1037/0022-006x.66.1.174

Thomas, M. L. (2011). The value of item response theory in clinical assessment: a review. *Assessment* 18, 291–307. doi: 10.1177/1073191110374797

Todorov, G. I., Mayilvahanan, K., Cain, C. K., and Cunha, C. (2018). *Screening Word Usage in People Affected by PTSD: An Unbiased, Cost Effective, and*

Novel Screening Method? *PsyArXiv* (Preprints). Available at: https://psyarxiv.com/y68fx/ (accessed October 10, 2019).

van den Berg, S. M., Paap, M. C. S., Derks, E. M., and Genetic Risk and Outcome of Psychosis (Group) investigators, (2013). Using multidimensional modeling to combine self-report symptoms with clinical judgment of schizotypy. *Psychiatry Res.* 206, 75–80. doi: 10.1016/j.psychres.2012.09.015

van der Linden, W. J., and Glas, C. A. (eds). (2000). *Computerized Adaptive Testing: Theory and Practice*. Dordrecht: Kluwer Academic.

van Groen, M. M., ten Klooster, P. M., Taal, E., van de Laar, M. A. F. J., and Glas, C. A. W. (2010). Application of the health assessment questionnaire disability index to various rheumatic diseases. *Qual. Life Res.* 19, 1255–1263. doi: 10.1007/s11136-010-9690-9

Weisscher, N., Glas, C. A. W., Vermeulen, M., and de Haan, R. J. (2010). The use of an item response theory-based disability item bank across diseases: accounting for differential item functioning. *J. Clin. Epidemiol.* 63, 543–549. doi: 10.1016/j.jclinepi.2009.07.016

Wohlfarth, T., van den Brink, W., Winkel, F. W., and ter Smitten, M. (2003). Screening for posttraumatic stress disorder: an evaluation of two self-report scales among crime victims. *Psychol. Assess.* 15, 101–109. doi: 10.1037/1040-3590.15.1.101

Wong, E., Ungvari, G. S., Leung, S. K., and Tang, W. K. (2007). Rating catatonia in patients with chronic schizophrenia: rasch analysis of the bush-francis catatonia rating scale. *Int. J. Methods Psychiatr. Res.* 16, 161–170. doi: 10.1002/mpr.224

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika* 56, 589–600. doi: 10.1007/bf02294492

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 21 November 2019
doi: 10.3389/fpsyg.2019.02461

Predictive Feature Generation and Selection Using Process Data From PISA Interactive Problem-Solving Items: An Application of Random Forests

Zhuangzhuang Han[1]*, *Qiwei He*[2]* *and Matthias von Davier*[3]*

[1] Teachers College, Columbia University, New York, NY, United States, [2] Educational Testing Service, Princeton, NJ, United States, [3] National Board of Medical Examiners, Philadelphia, PA, United States

OPEN ACCESS

Edited by:
Samuel Greiff,
University of Luxembourg,
Luxembourg

Reviewed by:
Timothy R. Brick,
Pennsylvania State University,
United States
Daniel W. Heck,
University of Marburg, Germany

***Correspondence:**
Zhuangzhuang Han
zh2198@tc.columbia.edu
Qiwei He
qhe@ets.org
Matthias von Davier
MvonDavier@nbme.org

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: 11 January 2019
Accepted: 17 October 2019
Published: 21 November 2019

Citation:
Han Z, He Q and von Davier M
(2019) Predictive Feature Generation
and Selection Using Process Data
From PISA Interactive
Problem-Solving Items: An
Application of Random Forests.
Front. Psychol. 10:2461.
doi: 10.3389/fpsyg.2019.02461

The Programme for International Student Assessment (PISA) introduced the measurement of problem-solving skills in the 2012 cycle. The items in this new domain employ scenario-based environments in terms of students interacting with computers. Process data collected from log files are a record of students' interactions with the testing platform. This study suggests a two-stage approach for generating features from process data and selecting the features that predict students' responses using a released problem-solving item—the Climate Control Task. The primary objectives of the study are (1) introducing an approach for generating features from the process data and using them to predict the response to this item, and (2) finding out which features have the most predictive value. To achieve these goals, a tree-based ensemble method, the random forest algorithm, is used to explore the association between response data and predictive features. Also, features can be ranked by importance in terms of predictive performance. This study can be considered as providing an alternative way to analyze process data having a pedagogical purpose.

Keywords: process data, interactive items, feature generation, feature selection, random forests, problem-solving, PISA

INTRODUCTION

Computer-based assessments (CBAs) are used for more than increasing construct validity (e.g., Sireci and Zenisky, 2006) and improving test design (e.g., van der Linden, 2005) through inclusion of adaptive features. They also provide new insights into behavioral processes related to task completion that cannot be easily observed using paper-based instruments (Goldhammer et al., 2013). In CBAs, a variety of timing and process data accompany test performance. This means that much more data from the response process of an answer is available in addition to correctness or incorrectness.

Along with assessing the core domains of Math, Reading, and Science, the Programme for International Student Assessment (PISA) introduced a problem-solving domain in the 2012 cycle, with fundamental technical support from computer delivery. It enabled interactive problems – problems in which exploration is required to uncover undisclosed information

(Ramalingam et al., 2014)—to be included in a large-scale international assessment for the first time (Organisation for Economic Co-operation and Development [OECD], 2014b). Dynamic records of actions generated during the item-response process form a distinct sequence representing participants' input and the internal state of the assessment platform. Analyzing these sequences can facilitate understanding of how individuals plan, evaluate, and select operations to achieve the problem-solving goal (e.g., Goldhammer et al., 2014; Hao et al., 2015; He and von Davier, 2016; Liao et al., 2019).

The problem-solving items in this new domain were typically designed as interactive tasks. The contents of these items cover a broad scope, from choosing an optimal geographic path between departure and destination points to purchasing metro tickets via a vending machine. Both the students' responses and the whole process of how students solved the problem in a sequence were captured in log files, such as clicking buttons, drawing lines, dragging on a scale, performing keystrokes to respond to open-ended items, and so on. The data contained in log files, referred to as *process data* in the present study, provide information beyond response data (i.e., whether the final response was correct or not).

While process data are expected to provide a broader range of information, the complex embedded structure demands an extension of existing analysis methods. These demands entail efforts to apply techniques used in other disciplines such as data mining, machine learning, natural language processing (NLP), social networking, and sequence data mining. These techniques serve two purposes: (1) creating predictive features/variables[1] associated with an outcome variable (i.e., feature generation) and (2) determining which features are the most predictive (i.e., feature selection).

The present study analyzed process data from a released PISA 2012 item (Organisation for Economic Co-operation and Development [OECD], 2014a)—Climate Control Task – that is intended to measure problem-solving skills of participants in science. The purpose of this study was twofold: first, to use process data obtained in a simulation-based environment to generate predictive features; and second, to identify the most important predictive features associated with success or failure on the task. The present study employed one of the tree-based ensemble methods – random forests – to select the most predictive features when considering students as the target of inferences.

The remainder of this paper is organized as follows. First, a brief overview of the methods is provided for generating features from process data and selecting important classifiers. The random forest algorithm is introduced and its potential use in analyzing process data is discussed. In the subsequent section, an integrated approach for generating features from process data and selecting features by the algorithm is introduced using a case study from the PISA 2012 problem-solving item. Results obtained from the introduced approach and their interpretations are then presented. Lastly, thoughts on the limitations and implications of the suggested approach are given.

OVERVIEW OF FEATURE GENERATION AND SELECTION USING PROCESS DATA

Generating Features Using Process Data

The principle of predictive feature generation is to maximize information exploration generated solely from timing and process data. This information may be indicative of respondents' problem-solving processes, which are associated with the problem-solving skills targeted in the assessment. As summarized in He et al. (2018), the features collected in log files can be roughly categorized into three groups: (1) behavioral indicators that represent respondents' problem-solving strategies and interactions with the computer, (2) sequences of actions and mini-actions that are directly extracted from test takers' process data, and (3) timing data such as total time on task, duration of respondent actions in the simulation environment, and time until first actions are taken by the respondent when solving a digital task.

Behavioral Indicators

Behavioral indicators are typically recorded at a higher, aggregated level. Although human-computer interactions are often accomplished through simple gestures or movements, in most cases, they are not automated actions but involve case-based reasoning and self-regulatory processes (Shapiro and Niederhauser, 2004; Azevedo, 2005; Lazonder and Rouet, 2008; Zimmerman, 2008; Brand-Gruwel et al., 2009; Bouchet et al., 2013; Winne and Baker, 2013). Therefore, to perform well on computer-based problem-solving tasks, one needs to have essential skills in using information and communication technology tools and higher-level skills in problem solving. Respondents have to decode and understand menu names or graphical icons in order to follow the appropriate chain of actions to reach a goal. Meanwhile, problem-solving tasks also require higher-order thinking, finding new solutions, and interacting with a dynamic environment (Mayer, 1994; Klieme, 2004; Mislevy et al., 2012; Goldhammer et al., 2014).

A typical example is the strategy indicator "vary one thing at a time (VOTAT)" studied in Greiff et al. (2015). This study showed that VOTAT was highly correlated with student performance. Note that solving complex, interactive tasks requires developing a plan consisting of a set of properly arranged subgoals and performing corresponding actions to attain the final goal. This differs from solving logical or mathematical problems, where complexity is determined by reasoning requirements but not primarily by the information that needs to be accessed and used (Goldhammer et al., 2013). In this sense, one could argue that the indicators of user actions should in some systematic way map onto the subgoals a user develops and applies to achieve a successful completion of the learning or assessment task.

Another example of a strategy indicator was derived from the problem-solving path and pace of examinees as studied in Lee and Haberman (2016). In this study, it was found that test takers adopted different strategies in solving reading tasks in an international language assessment and that these strategies were highly related to respondents' country, language, and cultural

[1] Predictor variables and covariates are also used interchangeably without being specifically mentioned in sections that follow.

background. For example, the typical strategy of test takers from two Asian countries was to skip the passage and view the questions first. Based on what the item's instructions requested, those test takers went back to read the passage and locate the information needed. Conversely, participants from two European countries were found to follow what was intended, that is, read the stimuli passage first and then answer the questions. These two strategies did not have a significant relationship to performance of test takers, although substantial performance differences and completion rates were found in the low-performing group.

Sequences of Actions From Process Data

The importance of sequence data in education has been recognized for decades. Agrawal and Srikant (1995) said "the primary task, as applied in a variety of domains including education, is to discover patterns that are found in many of the sequences in a dataset." Actions or mini-sequences that can be represented as *n*-grams (He and von Davier, 2015, 2016) are typical indicators to describe respondents' behavioral patterns. For instance, actions related to "cancel" (e.g., clicking on a cancel button in order to go back and change or check entries again) are sequence indicators, which are associated with test takers' cognitive processes and may indicate hesitation or uncertainty about next steps. Mini-sequences can not only show the actions adjacent to each other, but also the strategy link between the actions. For example, in He and von Davier (2016), strategy changes between the searching and sorting functions were successfully identified through analysis of bigrams and trigrams. More details on the use of *n*-grams for analyzing action sequences are given in the see section "Materials and Methods".

Some researchers have employed sequential pattern mining to inform student models for customizing learning to individual students (e.g., Corbett and Anderson, 1995; Amershi and Conati, 2009). Other researchers have employed sequential pattern mining to better understand groups' learning behaviors in designed conditions (e.g., Baker and Yacef, 2009; Zhou et al., 2010; Martinez et al., 2011; Anderson et al., 2013). As Kinnebrew et al. (2013) summarized, "identifying sequential patterns in learning activity data can be useful for discovering, understanding, and, ultimately, scaffolding student learning behaviors." Ideally, these patterns provide a basis for generating models and insights about how students learn, solve problems, and interact with the environment. Algorithms for mining sequential patterns generally associate some measures of frequency to rank identified patterns. The frequency of a pattern along the problem-solving process timeline can provide additional information for interpretation. Further, the observed variability across action-sequence patterns may play an important role in identifying behavioral patterns that occur only during a certain span of time or become more or less frequent than ones occurring frequently but uniformly over time, thus allowing us to explore what conditions lead to such changes (Kinnebrew et al., 2013).

Sequential pattern mining can be conducted via various approaches. For instance, Biswas et al. (2010) used hidden Markov models (HMMs; Rabiner, 1989; Fink, 2008) as a direct probabilistic representation of the internal states and strategies. This methodology facilitated identification, interpretation, and comparison of student learning behaviors at an aggregate level. As with students' mental processes, the states of an HMM are hidden, meaning they cannot be directly observed but produce observable output (e.g., actions in a learning environment).

As Fink (2008) pointed out, the development and spread in use of sequential models was closely related to the statistical modeling of texts as well as the restriction of possible sequences of word hypotheses in automatic speech recognition. Motivated by the methodologies and applications in NLP and text mining (e.g., He et al., 2012; Sukkarieh et al., 2012), a number of methods from these fields can be borrowed for application in process data analysis. For instance, the longest common subsequence introduced by Sukkarieh et al. (2012) to educational measurement for scoring computer-based Program for the International Assessment of Adult Competencies items was used in He et al. (2019) to extract the most likely strategy by respondent in each item by calculating the distance between individual sequences and reference ones. This approach allowed comparisons of respondents' behavior across multiple items in an assessment.

Features Generated From Timing Data

In addition to sequential data on actions taken by respondents during the problem-solving process, CBAs provide rich data on response latency or timing data. Each action log entry is associated not only with data on what a respondent did, but also when the action took place. These timestamps can be aggregated to an overall time measure for the survey, which is specific to the task, or measures that are specific to certain types of interactions such as keystrokes, navigation behavior, or time taken for reading instructions. Timing data at this level of resolution has led to renewed interest in how latency can be used in modeling response processes (e.g., DeMars, 2007; van der Linden et al., 2010; Weeks et al., 2016). In addition, timing data information is expected to be valuable in conjunction with the types of actions observed in the sequence data and to help us derive features that allow predicting cognitive outcomes such as test performance as well as background variables (Liao et al., 2019).

Predictive Feature Selection

Feature selection models play an essential role in identifying predictive indicators that can distinguish different groups, such as the correct and incorrect groups at the item level in problem-solving processes. A variety of models that have been developed in "big data" fields that relate to information retrieval, NLP, and data mining are also applicable to process data analysis. Here, we discuss some commonly used feature selection models that are popularly used in similar settings, ultimately focusing on one tree-based ensemble method – the random forest method – which will be further applied in this study.

As reviewed by Forman (2003) as well as Guyon and Elisseeff (2003), the feature selection approaches are essentially divided into wrappers, filters, and embedded methods. *Wrappers* utilize the learning machine of interest as a black box to score subsets of variables according to their predictive power. *Filters* select subsets of variables as a preprocessing step, independent of the chosen predictor. *Embedded* methods perform variable selection in the process of training and are usually specific to given learning machines. We introduced these three methods in details in the following subsections. In the embedded methods, the random forests method that has been used in this study is highlighted.

Wrapper Methods

These methods, popularized by Kohavi and John (1997), offer a simple and powerful way to address the problem of variable selection, regardless of the chosen machine learning approach. In their most general formulation, they consist of using the prediction performance of a given approach to assess the relative usefulness of subsets of variables. The wrapper methods that are most used in sequential forward selection or genetic search perform an exhaustive search over the space of all possible subsets of features, "repeatedly calling the induction algorithm as a subroutine to evaluate various subsets of features" (Guyon and Elisseeff, 2003). These methods are more practical for low-dimensional data but often are not for more complex large-scale problems due to intractable computations (Forman, 2003).

Filter Methods

These methods apply an intuitive approach in that the associations of each predictor variable with the response variable are individually evaluated, and those most associated with it are selected. For nominal response variables (the case considered in this study), measures of dispersion (also referred to as concentration or impurity depending on the context) such as Gini's impurity index and Shannon (1948)'s entropy are employed as the building blocks for measures of association between response variables and features (Haberman, 1982; Gilula and Haberman, 1995). In cases where response and features are both categorical, Goodman and Kruskal (1954) measure the association using the proportion of reduction of concentration if a predictor variable is involved. Other examples of measures of association can be found in, Theil (1970), Light and Margolin (1971), and Efron (1978).

Practices in area such as NLP implement an even more simplified approach by comparing the value of test statistics of association such as the chi-square statistic for the nominal response and categorical independent variable (Nigam et al., 2000; Oakes et al., 2001; He et al., 2012, 2014). Though some have raised concerns that this approach lacks statistical significance and soundness, its practical effectiveness in ordering the importance of categorical features makes it broadly accepted by certain audiences (Manning and Schütze, 1999; Forman, 2003). Applications can be founded in the recent literature about feature selection in large-scale assessment (He and von Davier, 2015, 2016; Liao et al., 2019).

Embedded Methods

These methods incorporate variable selection as part of the model training process. Compared with wrapper methods, they are more efficient and reach a faster solution by avoiding retraining a predictor from scratch for every variable subset investigated (Guyon and Elisseeff, 2003). For instance, the classification and regression tree (CART; Breiman et al., 1984) algorithm can be redesigned to serve this purpose. The random forest algorithm (Breiman, 2001), as an extension of CART that is a random ensemble of multiple trees, belongs to the family of embedded methods and is the method chosen for the current study. The random forest algorithm increasingly adjusts itself by randomly combining a predetermined number of single tree algorithms (shorten as trees in later sections). By aggregating the prediction results obtained from individual trees, the forest reduces prediction variance and improves overall prediction accuracy (Dietterich, 2000).

Some basic ideas about tree algorithms are reviewed here to facilitate understanding of the random forest algorithm. Let $X = X_1, \ldots, X_p$ for covariates and Y denote the outcome variable. Instead of establishing an analytical form of predicting Y from X, a decision tree grows by recursively splitting the space of covariates extended by the set X in a greedy way such that segments (nodes) created have the least impurity (for classification) or mean squared error (for regression) possible and are thus used to predict$Y$. Binary split – splitting a parent node into two child nodes – is conventionally employed and guided by the splitting rules. For classification, one of the rules is the Gini impurity index (Breiman et al., 1984; Breiman, 2001),

$$I_G(s, t) = 1 - \sum_k p_k^2(s, t),$$

where t denotes the current node, $p_k\,(s,\ t)$ is the frequency of class k in the samples of node t, and split s represents a certain numeric value or class label of a covariate X_j. If Y is binary, the above expression will be simplified as $1 - p_0^2\,(s,\ t) - p_1^2\,(s, t)$. It is intuitive that the index is a measure of dispersion: 1 indicates the utmost dispersion and 0 stands for the most extreme concentration. In other fields such as ecology, the index used to measure diversity is known as the Simpson-Gini Index due to its similarity to the Simpson Index (Peet, 1974). It should be noted that the estimate of $I_G\,(s, t)$ is biased for small samples if the sample frequencies $f_k\,(s,\ t) =\ n_k(s, t)/n(s, t)$ are directly used. This is because the unbiased estimate of $p_k^2\,(s, t)$ is $\frac{n_k(s,t)[1-n_k(s,t)]}{n(s,t)[1-n(s,t)]}$. A simple modification can be implemented to correct this bias.

The optimal split is determined by seeking the s that maximizes

$$\Delta I_G\,(s, t) =\ I_G\,(s, t) - \frac{1}{N_t}[N_{t_l} I_G\,(s, t_l) + N_{t_r} I_G(s, t_r)]$$

through the given predictors in set X. The quantity above indicates the decrease of impurity resulting from splitting the parent node t at s into the left child node t_l and the right child node t_r. Sample sizes (N_{t_l} and N_{t_r}) of child nodes are used to obtain the weighted impurity. For regression, the mean squared error is applied as the splitting rule (Breiman et al., 1984; Breiman, 2001).

Random forests ensemble individual decision trees through the following steps. First, subsets of samples are randomly drawn from the whole sample dataset and individual trees are grown based on each subset of samples. Note that data entries not chosen in each random draw are called "out of bag" data and kept for validating purposes. Second, for each individual decision tree in the random forest algorithm, it grows by recursively splitting a parent node into two or more child nodes with respect to a set of predictor variables as previously discussed. Rather than seeking the "best" cut point through all available predictor variables, the tree of random forests only examines through a set of *m* randomly chosen variables at each split. An individual tree stops to grow when a preset number of leaf nodes (nodes at the end of the tree that have no child nodes) or a threshold in terms of impurity of child nodes is reached. Third, final predicted responses are obtained by aggregating the prediction results over these fitted individual trees constructed using different subsets of covariates.

Even though the stability of an individual tree in terms of prediction is still not quite comparable with a typical logistic regression model fitted using all covariates, Breiman et al. (1984) argued that the variance is reduced because of the aggregation, which further enhances the overall prediction performance. Lin and Jeon (2006) showed that the random forest outperforms other less model-based predictive methods in cases with moderate sample sizes. In addition to the improvement on prediction performance, random forests also have other advantages in practice. As introduced above, only a certain number of covariates are selected to conduct each split when growing a decision tree. Such a feature allows the random forest algorithm to fit with a relatively larger number of predictor variables (especially for categorical variables) on a given sample size compared to other predictive methods such as linear models (e.g., generalized linear models), for which fitting with an extensive number of predictors may create data sparsity and reduce the numerical robustness.

In addition, two built-in variable selection methods of random forests, using two types of variable importance measures (VIMs)—(1) impurity importance and (2) permutation importance – have been successfully applied in fields such as gene expression and genome-wide association studies (Díaz-Uriarte and Alvarez de Andrés, 2006; Goldstein et al., 2011). The current study utilizes the permutation importance to select the most important variables extracted from the process data.

Impurity importance is quantified by accumulating $\Delta I_G(s, t)$ for each covariate over nodes of all trees. The accumulation is weighted by the sample sizes of nodes. While the importance measure enjoys all the computational convenience of the random forest algorithm, the splitting mechanism – just by chance – favors variables with many possible split points (e.g., categorical variables with many levels), resulting in a biased variable selection (Breiman et al., 1984; White and Liu, 1994). Much statistical literature further investigated this issue and proposed practical solutions (Kim and Loh, 2001; Hothorn et al., 2006; Strobl et al., 2007; Sandri and Zuccolotto, 2008). For instance, Strobl et al. (2007) reimplemented the random forest method based on Hothorn et al.'s (2006) conditional inference tree-structural algorithms (*ctrees*) to provide unbiased estimation of impurity importance. Instead of altering the algorithm, Sandri and Zuccolotto (2008) proposed a heuristic procedure to directly correct the bias of impurity measure by differentiating the "importance" resulting from characteristics of variables from the importance due to the association with the outcome variable.

As another built-in VIM of the random forest algorithm, the measure of permutation importance is free from this undesirable bias. Although it has been criticized for its computational inconvenience, the simple nature of the permutation importance measure becomes attractive as computation speed increases. The rationale of the permutation importance measure is as follows: First, a predictor variable, say X_j, is permutated in terms of the order of samples. Second, together with the other unaltered variables, another random forest algorithm is fit to compare with the algorithm constructed using unaltered samples. Permutation breaks the original association between X_j and Y, resulting in a drop of prediction accuracy for the testing data. Lastly, the rank of predictor variables can be established after applying this procedure to each covariate. In the present study, the permutation importance measure, also known as the mean decrease accuracy (Breiman, 2001), was implemented to conduct variable selection.

Tree-based ensemble algorithms also include bagging (Breiman, 1996) and boosting (Freund and Schapire, 1997). Bagging-tree algorithms are similar to random forests but are more straightforward in terms of randomizing the data and growing individual trees. Boosting-tree algorithms grow a sequence of single trees in a way that the latter grown tree fits the variation not explained by the former grown tree. Bayesian additive regression tree (BART; Chipman et al., 2010) is a tree ensemble method established in the Bayesian approach, offering a straightforward means of handling model selection by specifying a prior for the tuning parameter controlling the complexity of trees. Meanwhile, BART considers the uncertainty of parameter estimation with that of model selection. In addition, this method provides a flexible way to address the missing data issue by allowing for directly modeling the missing mechanism.

MATERIALS AND METHODS

Item Description and Data Processing

This study analyzed process data from a computer-based problem-solving item from PISA 2012 – Climate Control Task 1 (item code CP02501). The full-sample data has been made publicly available by the OECD[2]. The dataset for this item includes responses from 30,224 15-year-old in-school students from 42 countries and economies. Sample sizes of countries and economies are shown in **Table 1**.

This item (a snapshot of the item is shown in **Figure 1**) asked test takers to determine which of the three sliders controls temperature and which controls humidity, respectively. To obtain the correct answer, test takers were permitted to manipulate the sliders and monitor changes through the display. The answer to the task was given by drawing lines linking the diagrams to indicate the association between the inputs (sliders) and outputs. The correct solution is shown in **Figure 1**. The "reset" button undid previous simulations by clearing the display and resetting the sliders to their initial status. No limit was

[2] The dataset is available at http://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm.

TABLE 1 | Countries and economies with sample sizes.

Country and economy	Sample size
Australia	1,855
Austria	442
Belgium	726
Bulgaria	988
Canada	1,516
Chile	526
Chinese Taipei	494
Columbia	736
Croatia	962
Czechia	1,526
Denmark	636
Estonia	464
Finland	1,769
France	429
Germany	430
Hong Kong	433
Hungary	424
Ireland	407
Israel	440
Italy	453
Japan	1,005
Korean	449
Macao	519
Malaysia	938
Montenegro	917
Netherland	891
Norway	401
Poland	379
Portugal	486
Russia	504
Serbia	867
Shanghai-China	408
Singapore	469
Slovak	485
Slovenia	667
Spain	885
Sweden	418
Turkey	998
United Arab Emirates	1,023
United States	425
Uruguay	966

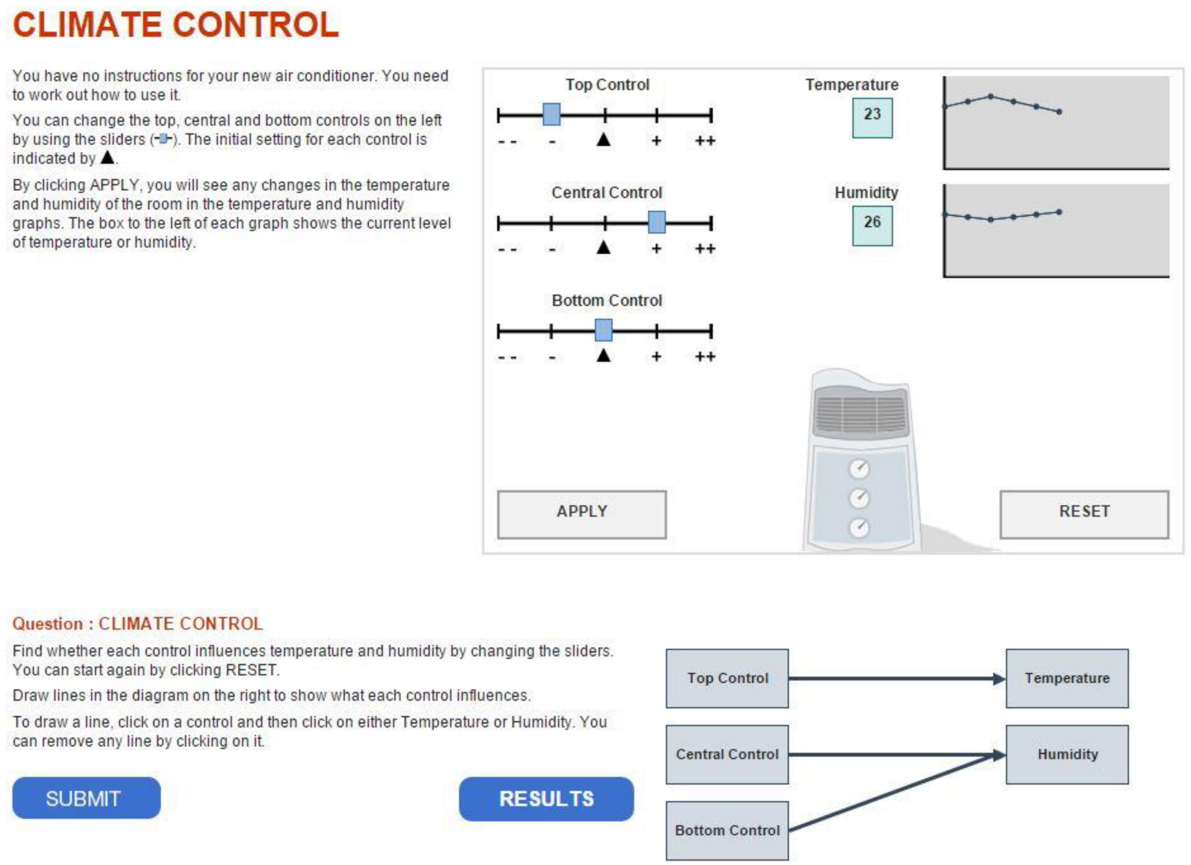
imposed on the number of steps of manipulation or rounds of exploration. Also, no time constraint was imposed on each item; however, the total test time of a test cluster (problem-solving items) was limited to 20 min. Either one or two clusters were randomly given to a participant depending on different assessment designs (Organisation for Economic Co-operation and Development [OECD], 2014b). The order of items in a cluster was fixed, and a former item could not be resumed once the next item had begun. According to different assignments of clusters, the position of Climate Control Task 1 varied across test forms. For this item, the average time spent by students was 125.5 s and the median time was 114.5 s; 95% of examinees spent from 22.2 s to 290.2 s on the item; only 1,149 participants (about 3.8% of the total sample) finished the task in 30 s or less, with a 5.1% rate of correctness. Given these results, later sections of the paper assume that the item is not considered as speeded for this sample in general and position effects, if any, are negligible. However, the analysis of the current study conducted without considering the speeded issue which should be noted as a limitation and further investigated by future research.

Items like Climate Control Task 1 are constructed using the MicroDYN approach (Greiff et al., 2012) that combines the use of the theoretical framework of linear structural equation models to systematically construct tasks (Funke, 2001) with multiple independent tasks to increase reliability. Briefly speaking, a system of causal relations (e.g., the first slider controls temperature) is embedded in a scenario that allows participants to explore input variables and observe the corresponding changes of output variables through a graphical representation. No specific prior domain knowledge is required for this type of task in general. However, examinees need to gain and have command of the knowledge by exploring and experimenting before providing appropriate answers. For such tasks, a strategic knowledge for effective exploration is crucially important (Greiff et al., 2015)—that is, the VOTAT (vary one thing at a time; Tschirgi, 1980) strategy; this term is also known as the control-of-variable strategy (Chen and Klahr, 1999) in developmental psychology.

In PISA 2012, a partial credit assignment – 0 for incorrect, 1 for partially correct, and 2 for correct – was used to score the responses of Climate Control Task 1. Partial credit was given if a student explored the simulation by using the VOTAT strategy efficiently – only varying one control at a time when trying to change the status of each control individually at least once, regardless of actions being in adjacent attempts or in a round before resetting – but failed to correctly represent the association in a diagram.

To show that the VOTAT strategy is strongly related to performance on the item, Greiff et al. (2015) restricted polytomous responses as dichotomous by treating partially correct as incorrect and then investigated the association between the dichotomous responses with the indicator of applying the VOTAT strategy efficiently alongside other covariates. Following the same settings, the present study explored the association between the binary responses and the indicator of the use of the VOTAT strategy together with other covariates created from the process data to find out (1) whether the current partial scoring rubric was still supported by the prediction model (i.e., random forests)—namely, whether the VOTAT variable was still the most associated factor with responses while interacting with other covariates – and (2) whether the rubric was still sufficient compared with the new predictor features extracted from the process data. It should be noted that the restriction of response variable may not be applicable for items that are intended to measure a construct other than the interactive complex problem-solving (Cheng and Holyoak, 1985; Funke, 2001) skills or constructed without using the MicroDYN approach.

Table 2 shows a section of the postprocessed log file—that is, a readable process dataset whose entries are actions



CLIMATE CONTROL

You have no instructions for your new air conditioner. You need to work out how to use it.

You can change the top, central and bottom controls on the left by using the sliders (-▪-). The initial setting for each control is indicated by ▲.

By clicking APPLY, you will see any changes in the temperature and humidity of the room in the temperature and humidity graphs. The box to the left of each graph shows the current level of temperature or humidity.

Question : CLIMATE CONTROL

Find whether each control influences temperature and humidity by changing the sliders. You can start again by clicking RESET.

Draw lines in the diagram on the right to show what each control influences.

To draw a line, click on a control and then click on either Temperature or Humidity. You can remove any line by clicking on it.

FIGURE 1 | A snapshot of the problem-solving item Climate Control in PISA 2012.

listed in chronological order. The even number indicates the actions belong to a certain test taker. The type of action, as well as the corresponding timestamp, was recorded for each action. Among the action types, "apply" represents actions related to manipulation of sliders because, after setting sliders, a test taker needed to hit the "apply" box, as shown in **Figure 1**, to see the changed value of temperature and humidity displayed. The changed status of sliders was recorded in the columns "top slider," "central slider," and "bottom slider." The value of status ranges from −2 to 2. Similarly, the action type "diagram" represents drawing a line to link diagrams, as shown at the bottom right of **Figure 1**. The six-digit binary string shown in the table was used to record the association among diagrams that has been established. For example, "100101" indicates that the top slider controls temperature, whereas the central and bottom sliders control humidity.

To facilitate the analysis, observed sequences of actions were collapsed into respective strings. To obtain such a string, each type of action is abbreviated using a single capital letter: "S" for "start," "E" for "end," "R" for "reset," "A" for "apply," and "D" for "diagram." It should be noted that consecutive "D" actions were collapsed into a single "D" action because information related to drawing lines to connect the diagrams is not of central interest in the present study. For the sequence of actions shown in **Table 2**, it can be simplified as "SRAAAAARDE."

Feature Generation

In this study, features (predictor variables) extracted from the process data can be summarized in three categories: variables extracted from action sequences using *n*-gram methods, behavior indicators, and time-related variables.

N-gram methods decode a sequence of actions into mini-sequences (e.g., a string of *n* letters in length where the letters remain in the same order as the original sequence of actions) and document the number of occurrences of each mini-sequence. Unigrams, analogous to the language sequences in NLP, are defined as "bags of actions," where each single action in a sequence collection represents a distinct feature. However, unigrams are not informative in term of transitions between actions. Bigrams and trigrams are considered in this study, with action sequences broken down into mini-sequences containing two and three ordered adjacent actions. Note that the *n*-gram method is productive in creating features based on sequence data without loss of much information about the order of sequence. This class of methods has become widely accepted for feature engineering in fields such as NLP and genomic sequence analysis and was recently applied to analyze

TABLE 2 | An example of process data for a test taker solving the climate control item.

Event	Time	Event_order	Event_ type	Top_ slider	Central_slider	Bottom_slider	Temp_value	Humid_value	Diag_state
START_ITEM	1288.1	1	start	NULL	NULL	NULL	NULL	NULL	NULL
ACER_EVENT	1291.9	2	reset	0	0	0	25	25	NULL
ACER_EVENT	1338.4	3	apply	1	1	1	27	28	NULL
ACER_EVENT	1346.8	4	apply	1	1	2	29	33	NULL
ACER_EVENT	1350.1	5	apply	1	2	2	31	36	NULL
ACER_EVENT	1354.5	6	apply	2	2	2	35	36	NULL
ACER_EVENT	1361.1	7	apply	2	1	1	36	36	NULL
ACER_EVENT	1361.1	8	reset	0	0	0	25	25	NULL
ACER_EVENT	1375.3	9	diagram	NULL	NULL	NULL	NULL	NULL	000000
ACER_EVENT	1376.2	10	diagram	NULL	NULL	NULL	NULL	NULL	000000
ACER_EVENT	1400.1	11	diagram	NULL	NULL	NULL	NULL	NULL	000000
ACER_EVENT	1402.1	12	diagram	NULL	NULL	NULL	NULL	NULL	000001
ACER_EVENT	1406.8	13	diagram	NULL	NULL	NULL	NULL	NULL	000001
ACER_EVENT	1408.4	14	diagram	NULL	NULL	NULL	NULL	NULL	000101
ACER_EVENT	1410.2	15	diagram	NULL	NULL	NULL	NULL	NULL	000101
ACER_EVENT	1410.6	16	diagram	NULL	NULL	NULL	NULL	NULL	100101
END_ITEM	1416.1	17	end	NULL	NULL	NULL	NULL	NULL	NULL

"Event" and "event_type" indicate the type of the current action. "Time" and "event_num" show the time point and order of the current action. "Top_slider," "central_ slider," and "bottom_ slider" provide information about the status of each control. "Temp_value" and "humid_value" give the simulated results. "diag_state" gives information on the linking among diagrams. Each type of event is abbreviated using a single capital letter: "S" for "start," "E" for "end," "R" for "reset," "A" for "apply," and "D" for "diagram." Data source: This table is extracted from "Log-file databases for released PISA 2012 computer-based items data for problem solving" at http://www.oecd.org/pisa/pisaproducts/database-cbapisa2012.htm.

process data in large-scale assessment (He and von Davier, 2015, 2016). For example, an *n*-gram can break the action string "SRAAAAARDE" into "S(1), A(5), R(2), D(1), E(1)" for unigrams, "SA(1), AR(1), AA(4), RA(1), RD(1), DE(1)" for bigrams, and "SRA(1), RAA(1), AAA(3), AAR(1), ARD(1), RDE(1)" for trigrams, where the numerals in brackets represent the number of occurrences.

Behavior indicators can also be generated from sequences of actions. Changes to input variables (the positions of controls) shed light on participants' problem-solving strategies and behaviors. As discussed earlier, partial credit was given to students who explored the connection between the inputs and outputs by utilizing the VOTAT (vary one thing at a time) strategy across the three controls at least once. Greiff et al. (2015) treated this scoring rubric as an indicator variable (i.e., VOTAT) and showed that it was highly associated with the probability of answering this item correctly and overall performance on the test.

This study created an ordinal categorical variable with four levels – from 0 to 3 – each number indicating on how many controls a student has used the VOTAT strategy. This ordinal variable was referred to as "VOTAT group" in the analysis. Another variable named "VOTAT num" was created to count the number of times that a student used the VOTAT strategy regardless of which control he or she applied the strategy to. Additionally, the order of "A" and "D" in a sequence of actions could convey information about examinees' decisiveness or hesitancy of their decision-making process. For example, the action string "SRAAAAARDE" can be categorized as a meta-strategy "AD sequence," implying the examinee "draws" the diagrams right after "applying" the simulations on sliders rather than jumping back and forth between applying sliders and drawing diagram lines.

Table 3 shows the distribution of the AD sequence variable, where N indicates the cases in which participants did not conduct an experiment or generate diagrams. Note that the AD sequence's having an undue number of levels not just hindered interpretation but also caused data sparsity in analysis that followed. Thus a "compact" version of AD sequence with fewer levels was created as shown in **Table 4**. **Figure 2** illustrates how to create the contracted levels in **Table 4** by a tree-like diagram.

Process data also provide rich information related to time. Process data includes timestamps of actions, allowing the time spent on a specific action to be calculated by taking the difference of the time of two adjacent actions. Several time-related predictor variables can be generated as follows. "A time" and "D time" indicate the accumulated time spent on manipulating controls and drawing diagrams, respectively. For example, for an action sequence "SADRE," "A time" is the time used after hitting the "start" box and before hitting the "apply" box; "D time" is the time spent after hitting the "apply" box and before drawing a line among diagrams. By a similar token, "E time" records the time spent after conducting the last action before hitting the "end" box. A special case is "R time," which represents the time spent after hitting the "reset" box but before conducting the next Action. "time_bf_action" records the time span between "start" and the first action after "start," which can be considered as the time spent on reading and perceiving the task.

Given the feature generation method described above, 77 variables were created from the process data (a snapshot of the process data is presented as **Table 2**), as presented in **Table 5**. Note that time-related features in this study were binned with

TABLE 3 | All levels of AD sequence with sample size and percentage of correctness.

AD Behavior	Total	Correct	Percentage (%)
AD	6490	2377	36.63
ADA	1118	522	46.69
ADAD	2996	1648	55.01
ADADA	697	401	57.53
ADADAD	8004	6470	80.83
ADADADA	1648	1459	88.53
ADADADAD	777	558	71.81
ADADADADA	250	188	75.20
ADADADADAD	167	115	68.86
ADADADADADA	64	41	64.06
ADADADADADAD	74	53	71.62
ADADADADADADA	29	17	58.62
ADADADADADADAD	15	8	53.33
ADADADADADADADA	8	6	75.00
ADADADADADADADAD	6	2	33.33
ADADADADADADADADA	7	3	42.86
ADADADADADADADADAD	4	1	25.00
ADADADADADADADADADA	3	1	33.33
ADADADADADADADADADADAD	1	0	0.00
ADADADADADADADADADADADA	1	1	100.00
ADADADADADADADADADADADADA	1	0	0.00
ADADADADADADADADADADADADADA	1	1	100.00
ADADADADADADADADADADADADADAD	1	1	100.00
DA	803	123	15.32
DAD	398	137	34.42
DADA	232	74	31.90
DADAD	190	91	47.89
DADADA	108	40	37.04
DADADAD	345	259	75.07
DADADADA	124	76	61.29
DADADADAD	84	54	64.29
DADADADADA	38	18	47.37
DADADADADAD	22	11	50.00
DADADADADADA	27	7	25.93
DADADADADADAD	11	5	45.45
DADADADADADADA	10	0	0.00
DADADADADADADAD	10	7	70.00
DADADADADADADADA	12	2	16.67
DADADADADADADADAD	6	2	33.33
DADADADADADADADADA	8	0	0.00
DADADADADADADADADADA	3	2	66.67
DADADADADADADADADADAD	1	0	0.00
DADADADADADADADADADADA	3	2	66.67
DADADADADADADADADADADAD	2	1	50.00
DADADADADADADADADADADADA	6	3	50.00
DADADADADADADADADADADADAD	1	0	0.00
DADADADADADADADADADADADADAD	1	1	100.00
DADADADADADADADADADADADADADADA DADADA	3	0	0.00
N	5414	267	4.93

TABLE 4 | All contracted levels of AD sequence with sample size and percentage of correctness.

	Total	Correct	Percentage (%)
Incomplete	5414	267	4.93
Start from D	2448	915	37.38
AD only	6490	2377	36.63
1<=AD<3	4811	2571	53.44
AD>=3	11061	8925	80.69

equal percentiles in terms of their frequencies – the frequency of each bin ranges from 10 to 25% of the sample depending on the variables. This was done essentially due to the nature of the tree models: continuous variables are discretized to find the best "split" point, as discussed in previous sections. This inherent discretization mechanism tends to create data sparsity when the distribution of a continuous variable is "discontinued" (i.e., having extreme low density at the area between modes), which increases the chance of encountering a computation failure. Therefore, to reduce this chance, practitioners "stabilize" the distributions of these "discontinued" variables by binning before feeding the variables to fit the algorithm. In this study, binning was also applied to *n*-gram features with levels having sparse sample sizes. However, it should be noted that binning entails a risk of losing information about these variables.

Feature Selection

Feature selection was conducted using the R package *randomForest* (Liaw and Wiener, 2002). The selection began with seeking the random forest algorithm having the optimal complexity to fit the dataset. In this study, the complexity of the random forest algorithm is characterized by combinations of number of trees (*ntree*) and number of predictor variables used to grow a tree (*mtry*). Empirical studies (Breiman, 2001; Mitchell, 2011; Janitza and Hornung, 2018) showed that *mtry* and *ntree* are more influential than other factors in controlling the complexity of the random forest algorithm. In this study the size of a tree (i.e., the number of generations or the total number of nodes) was not restricted and the number of branches used at each split was fixed at 2. The present study was focused on exploring the combinations of *mtry* and *ntree,* where *ntree* = 100, 300, 500, and *mtry* = 4, 6, 8, 10, 12.

Cross-Validation

A typical way to find the optimal model complexity (i.e., the combination of tuning parameters) is to compare the fitted models by their validation error. The validation error is obtained by holding out a subset of the sample (validation set), using the retained sample (training set) to fit the classification algorithm, and then estimating the prediction error by applying the fitted algorithm to the validation set. To efficiently utilize data with a limited size, practitioners (Breiman and Spector, 1992; Kohavi, 1995) have recommended five- or ten-fold cross-validation. In the case of five-fold cross-validation, the data is split into five roughly equal parts. A loop of validations is then conducted – each part is labeled as the validation set once to estimate the prediction error of the random forest model fitted using the other four parts. In a data-rich situation, Hastie et al. (2009) recommended to isolate an additional set (the test set) from

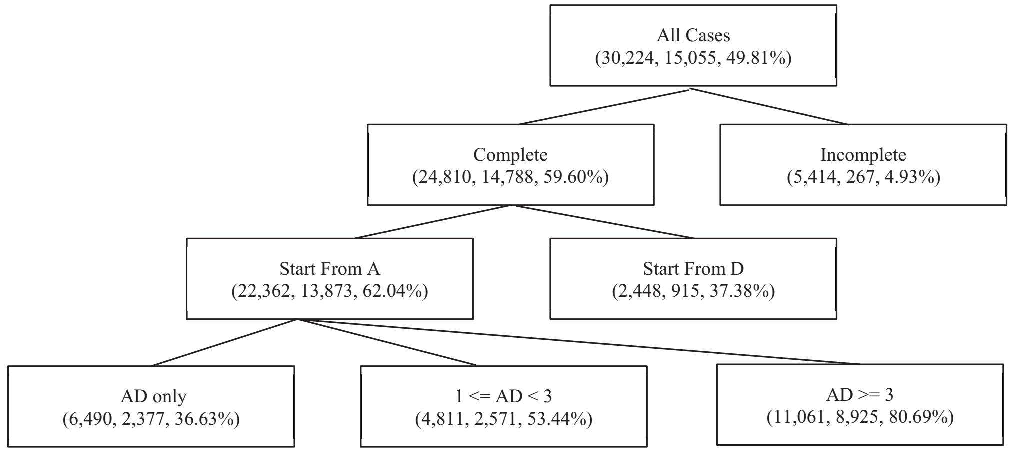


FIGURE 2 | A tree-based diagram for contracted levels of the AD sequence. Indices in parentheses are sample size, number of correct responses, and conditional probability of correctness, respectively, for each class or contracted class of the "AD sequence" variable.

the sample before conducting cross-validation. This set is used to compute the prediction error for the final chosen model. It can also be considered as an assessment of the generalization performance of the chosen model on independent data. The present study randomly selected roughly 10% of the sample (3,000 students) as the test set; the rest was separated into five folds for the training-validation purposes.

Variable Selection and Backward Elimination

The core idea of validation is to keep the validation sample from being "seen" by the model training process. Such a principle must also be obeyed when variable selection is involved. An example of violating this rule would be to conduct variable selection based on the whole sample before tuning model parameters based on cross-validation (Hastie et al., 2009).

The variable selection implemented in the current study is based on the recursive feature elimination in Guyon et al. (2002) that iteratively rules out features at the lower end of the ranking criterion. Together with random forests, recursive feature elimination has been successfully employed in genome-wide association studies (e.g., Jiang et al., 2009). The variable selection approach suggested in the present study is not just an application of recursive feature elimination using the random forest algorithm with a specific focus on the process data, but a modification with an emphasis on end-to-end cross-validation.

TABLE 5 | Variables generated from process data of climate control task 1.

	Total	Generated Features
Unigram	3	D, R, A
Bigram	16	DD, AA, RA, AR, AD, DA, AE, SD, SA, DR, DE, RD, RE, RR, SR, SE
Trigram	48	ADD, AAR, SRD, DDR, AAE, DRE, AAA, ARD, SDR, ADE, RAA, RRE, DDD, DAR, ARR, DAA, RDA, RRA, DAD, SDA, RRR, AAD, RAD, RRD, ADR, ARE, DRR, RDE, DRR, SRA, ADA, SAR, SRE, ARA, RAR, SDE, DRA, RDD, RDR, SDD, DAE, SAR, DDA, DRD, SRR, SAA, SAD, RAE
Behavioral indicators	4	AD sequence, VOTAT group, VOTAT num, n_actions
Time-related features	6	D time, A time, R time, E time, total time, time_bf_action.
Total	77	

Box 1 outlines the suggested backward elimination algorithm for variable selection. Note that to prevent variable selection (i.e., ranking) from seeing the data used for model training (i.e., parameter tuning in this study), the training-validation dataset was divided into five disjoint subsets in this recursive selection process so that at each backward elimination parameter tuning can be conducted using four of the subsets of data while variable ranking can be performed separately based on the other subset. This suggested approach follows the principles of variable selection for study design recommended by Brick et al. (2017).

As indicated by **Box 1**, the backward elimination also documents how the validation performance of the fitted model changes as the number of features reduces, which was illustrated in **Figure 3**. The number of selected features was decided by drawing a cutoff line around where the first large drop in

BOX 1 | Backward elimination algorithm for feature selection.

```
randomly split the training-validation dataset into 5 disjoint subsets.
$X_1, \ldots, X_5$ are sets of covariates; they are all same with 77 covariates at the beginning;
repeat the followings until the covariate sets $X_1, \ldots, X_5$ are empty:
    for $k$ in $\{1, 2, \ldots, 5\}$:
    hold the $k$-th dataset out for ranking;
        for each combination of mtry and ntree:
            conduct a five-fold cross-validation using the other 4 datasets and covariates left in $X_k$;
            obtain cross-validated prediction error $e$ for the current combination;
    find the optimal mtry and ntree by comparing $e$ across all combinations;
    fit a random forest using the $k$-th dataset and the optimal parameters;
    obtain the importance rank and remove the least important feature from $X_k$.
end
```

prediction performance begins (i.e., 49 in **Figure 3**). Setting this cutoff line here is like selecting the number of factors using the scree plot (Cattell, 1966). Given this threshold number (i.e., $77 - 49 = 28$), five sets with 28 selected features were obtained, and their intersection gives the final selected set of features (26 features).

The backward elimination in **Box 1** has five separated iterative variable ranking processes, which could be somehow regarded as an implicit self-validation. However, the determination of the cut-off line shared by the five ranking processes (i.e., the feature screening) should be further validated if data are rich enough. Instead of having one training-validation set, five disjointed training-validation sets (notice this is different from the five shown in **Box 1**) were established after the test set was held out. Backward elimination shown in **Box 1** was conducted for each of the five sets. Accordingly, five sets of final selected features were obtained. **Table 6** shows the intersection of these five sets of selected features.

The backward elimination in **Box 1** was structured using a nested loop that might cause inefficiency. Practitioners can increase the number of features eliminated for each round to reduce computation burden. Plus, as noted by Breiman (2001), the value of *mtry* set around the square root of the number of predictors seems to have minimal effect on validation performance; to increase computational efficiency, one can utilize this deterministic way to adapt the value of *mtry*. In addition, to further increase algorithmic efficiency, researchers (Breiman, 2001; Nicodemus and Malley, 2009; Zhang et al., 2010; Goldstein et al., 2011; Oliveira et al., 2012) recommended employing out-of-bag error as an alternative to cross-validation error. Simulation studies (Mitchell, 2011; Janitza and Hornung, 2018) showed that although out-of-bag error tends to overestimate true error rate when "$n<<p$"—that is, the sample size is far less than the number of predictors, the overall validation performance is not substantially affected by means of out-of-bag error to determine model complexity. The present study also performed a backward elimination boosted by using the above suggestions, which obtained consistent results with the plain approach shown in **Box 1** in terms of variable selection. Such results were not presented in the manuscript for the sake of simplicity.

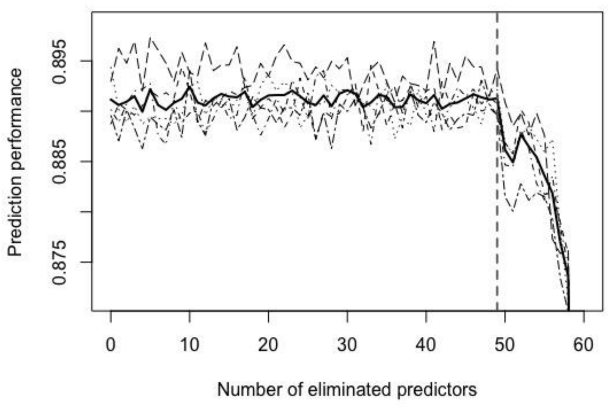


FIGURE 3 | Prediction performance versus number of eliminated predictors for a backward elimination. Dashed lines record the change of validation performance (classification accuracy) for each training set as the number of eliminated feature increases; the bold solid line represents the average performance for five-fold; the vertical dashed line (the number of excluded features=49) indicates where a large reduction of prediction performance begins.

TABLE 6 | Features selected through the five-fold validated backward elimination.

21 features	D, AD sequence, VOTAT num, **DD**, DDD, VOTAT group, **DDE**, RA, **AD**, R, D time, R time, n_actions, A, AAA, **ADD**, **AR**, **DA**, ADR, **DRA**, **DR**.

Boldfaced cases indicate features considered redundant. Such features are removed from the set of selected features for analysis that follows.

RESULTS

The final set of selected features includes ordinal and binary categorical variables. Pairwise associations among these ordinal variables were measured using the Goodman-Kruskal gamma (γ; Goodman and Kruskal, 1954) with value from -1 (discordant) to 1 (concordant). Given the measure, the final set can be further reduced by removing the redundant features highly related to others.

Among all pairs, "DD" was highly associated with "DDD"($\gamma = 0.76$); "AR" and "RA" was associated with $\gamma = 0.71$; other well-associated pairs ($\gamma > 0.6$) included "AD sequence" with "AD," "AD sequence" with "DA," "AD" with "ADD," "DRA" with "ADR," "DRA" with "DR," and "DD" with "DDE."[3] It is not surprising that "AD sequence" was highly correlated with "AD" and "DA." "AD sequence" was preferred since it covered more information than "AD" and "DA" do, as discussed earlier. "DDD" was greatly associated with "DD;" trigram was preferable in this case since it contained more detailed information. "DDE" conveyed trivial information compared to "DD" and "DDD," as did "ADD" to "AD." "AR" and "RA" covered similar information, as did "DRA" with "DR" and "ADR;" the one with higher rank of permutation importance was preferred. In sum, eight features (boldfaced in **Table 6**) were excluded: "AD," "DA," "ADD," "DDE," "DRA," "AR," "DD," and "DR."

With the 13 remaining features, a random forest was fitted with the parameter set where *ntree* = 100 and *mtry* = 4. The parameter combination was chosen based on validation performance of the test set that had been held out at the beginning. Applying the test set here was necessary since the association measured above was based on the entire validation-training sample, which means that variables selected using γ had already "seen" the validation data. Similarly, another random forest was fitted with 77 features; the parameter set was tuned using the test data, where *ntree* = 300 and *mtry* = 9. Here the Goodman-Kruskal tau (τ; Goodman and Kruskal, 1954) was used

[3] As a reminder, "D" refers to drawing the diagram, "A" to applying the simulations on the slider, "S" to start, and "R" to reset.

to measure the proportional reduction of incorrect prediction for the full and the reduced model, respectively, with regard to the random guess based on observed distribution of responses, where $\tau_{77} = 0.810$ and $\tau_{13} = 0.797$. In this regard, the reduced model performed decently in comparison to the full model.

Features of the simple model ranked by the permutation importance measure are shown in **Table 7**. Unigram "D," "R," and "A" ranked high in the list since they are basic elements constituting action sequences. Furthermore, "D" and "R" are not just fundamental but also imply a student's decisiveness. Using only a few necessary steps of drawing arrows or applying the reset function only a limited number of times might indicate confidence in providing a correct solution. "VOTAT group" and "VOTAT num" are both critical as shown in the list, which is consistent with the results found by Greiff et al. (2015). The top-ranked "AD sequence" indicates that contracting levels shown in **Figure 2** work fine in summarizing students' behaviors on experimenting. Grams such as "AAA," "ADR," and "RA" offer interesting perspectives. For instance, students having a large number of "AAA" tended to show certain patterns in their actions: drawing diagrams right after applying experiments (i.e., the level "AD only" in the feature "AD sequence") and applying the VOTAT scheme across the three sliders. In further investigating these students, we found that they attempted to create an increasing or decreasing slope of the value of temperature or humidity in the display by repeatedly hitting the "apply" box while fixing the sliders at one particular status, indicating a relatively sophisticated behavior of solving the problem. Frequent usage of "ADR" and "RA" indicated participants utilized the reset function to assist their experimenting and exploration on inputs. "D time" and "R time" can be regarded as time spent on deliberation.

DISCUSSION

The aim of the present study is to pedagogically suggest an integrated approach to analyze action sequences and other information extracted from process data. Feature generation and selection are two essential parts of the suggested approach and should be treated with equal importance. Features in this study were created following both top-down and bottom-up schemes. The former generates features based on hypotheses that might be developed by item designers and content experts. The latter, as an example, extracts features by utilizing *n*-gram methods and related methods breaking up the action sequences. Thus, *n*-gram translates the action sequences into mini-sequences along with their frequencies. Features generated by both schemes are presented in the final set of selected predictive features. The random forest algorithm was implemented in the feature selection part, which simultaneously handled (1) a massive number of categorical predictor variables, (2) the complexity of the variable structure, and (3) model/variable selection in a computationally efficient way. The utility of the suggested approach has been illustrated by implementing it in a publicly available dataset.

TABLE 7 | Features ranked by permutation importance measure (mean decrease accuracy).

Feature	Mean decrease accuracy
D	0.199
VOTAT group	0.056
AD sequence	0.042
VOTAT num	0.023
R	0.022
R time	0.018
DDD	0.017
n_actions	0.015
RA	0.014
A	0.013
D time	0.009
AAA	0.008
ADR	0.007

The suggested approach is not free from limitations. First, the feature generation process involves breaking up action sequences into mini-sequences encoded as *n*-grams, suggesting that the information contained in the order of the action sequences – for example, the "longer term" dependencies among actions – would not be completely preserved and exploited. As an outcome, only limited amounts of behavioral indicators are generated; information embedded in students' action sequences might not be fully utilized. For example, the range of states of controls explored by a student is a variable likely associated with the response variable. Technically speaking, to preserve more "complete" information when analyzing action sequences, sequence-mining approaches (e.g., SPADE; Zaki, 2001) employed to find common subsequences provide a possible alternative. Also, ideas stemming from cognitive and learning studies offer a theoretical basis of creating features from action sequences; for example, some studies (Jiang et al., 2015, 2018) employed sequential pattern mining to analyze learning skills and performance in immersive virtual environments.

Second, most features, if not all, are ordinal categorical variables representing frequency. As noted in the previous section, some variables present in excessive levels could cause an issue of data sparsity when conducting the random forest algorithm. This study used equal-percentile binning to address this issue at the expense of losing information provided by the original variables. The sensitivity of binning needs to be further investigated.

Third, the CART-based random forest algorithm using the Gini-impurity index to split nodes (e.g., the *randomForest* R package used in this study) implemented in this study is generally a suboptimal choice. Strobl et al. (2007) showed that the algorithm tends to favor categorical variables with extensive levels as well as a cluster of variables that are highly correlated. The modified random forest algorithm proposed by Strobl et al. (2007) using the conditional inference tree introduced by Hothorn et al. (2006) should be explored in the context of process data for future studies.

Fourth, even though the efficiency of the suggested backward elimination can be increased by using several steps noted in the previous section, the computation burden is still a concern for the present study. Backward elimination with the specifications shown in **Box 1** was validated using a five-fold dataset, which took about 19,872 s in total on a Mac Pro desktop with a 3.5 GHz CPU and 16 GB of RAM.

Fifth, like other data-driven algorithms, the random forest approach is not straightforward regarding model interpretation. For example, hypothesis tests on marginal effects of features are not sustained in random forests; the directions of marginal effects are not directly accessible, either. Friedman (2001) suggested plotting the partial dependence between the feature and the outcome variable (logit is used if the outcome variable is categorical) to access the marginal effects. This display method has been implemented in the R package *randomForest* as the function partialPlot. It is sensible to apply models with more restricted functional forms, such as linear models, to conduct an *ad hoc* analysis based on the selected features.

Sixth, the random forest algorithm is a data-driven method that is sensitive to sample characteristics. Meanwhile, PISA is an international large-scale assessment involving mixed-type forms of tests and multistage sampling designs. The question on how the sampling designs affect the analysis using data-driven methods (i.e., random forests) in terms of estimation stability is beyond the scope of this study. It is appealing that future methodological research could provide guidance concerning the correct use of cross-validation in different test designs.

Last, the exploratory nature of the suggested approach comes with the purpose of the study. Although interesting patterns of behaviors have been found by the suggested approach, it is still difficult to test a cognitive or psychometric theory with it.

The suggested method offers an alternative to the generation and selection of informative features from a massive amount of process data, given the increasing attention to exploring the usage of process data along with response data in large-scale assessments. Generalizability of the method can be explored by applying it to multiple tasks constructed using a similar approach such as MicroDYN and comparing it with other variable-selection approaches in terms of practical significance.

ETHICS STATEMENT

This study is a secondary analysis based on released datasets from PISA 2012 log data files (http://www.oecd.org/pisa/pisaproducts/database-cbapisa2012.htm). No additional data were collected from human subjects for this particular study.

AUTHOR CONTRIBUTIONS

ZH contributed to the development of methodology exploration, model estimation procedures, conduction of the data analysis, and drafting and revision of the manuscript. QH contributed to the development of the methodological framework, supervision on the model estimation procedures, conduction of the data analysis, and drafting and revision of the manuscript. MD contributed to providing suggestions on the methodological framework and the model estimation procedures, and reviewing and revision of the manuscript.

FUNDING

This study was partially supported by the Internship Program under Educational Testing Service and the National Science Foundation (NSF – Award #1633353).

ACKNOWLEDGMENTS

The authors would like to thank Larry Hanover for his help in editing the manuscript and Dr. Xiang Liu for his valuable suggestions.

REFERENCES

Agrawal, R., and Srikant, R. (1995). "Mining sequential patterns," in *Proceedings of the Eleventh IEEE International Conference on Data Engineering*, Taipei.

Amershi, S., and Conati, C. (2009). Combining unsupervised and supervised classification to build user models for exploratory learning environments. *J. Educ. Data Min.* 1, 18–81.

Anderson, E., Gulwani, S., and Popovic, Z. (2013). "A trace-based framework for analyzing and synthesizing educational progressions," in *Proceedings of the Special Interest Group on Computer-Human Interaction (SIGCHI) Conference on Human Factors in Computing Systems*, (Paris).

Azevedo, R. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educ. Psychol.* 40, 199–209.

Baker, R., and Yacef, K. (2009). The state of educational data mining in 2009: a review and future visions. *J. Educ. Data Min.* 1, 3–16.

Biswas, G., Jeong, H., Kinnebrew, J., Sulcer, B., and Roscoe, R. (2010). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Res. Pract. Technol. Enhanc. Learn.* 5, 123–152. doi: 10.1142/S1793206810000839

Bouchet, F., Harley, J. M., Trevors, G. J., and Azevedo, R. (2013). Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning. *J. Educ. Data Min.* 5, 104–146.

Brand-Gruwel, S., Wopereis, I., and Walraven, A. (2009). A descriptive model of information problem-solving while using internet. *Comput. Educ.* 53, 1207–1217. doi: 10.1016/j.compedu.2009.06.004

Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 24, 123–140. doi: 10.1007/BF00058655

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. I. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Breiman, L., and Spector, P. (1992). Submodel selection and evaluation in regression. *Int. Statist. Rev.* 60, 291–319.

Brick, T. R., Koffer, R. E., Gerstorf, D., and Ram, N. (2017). Feature selection methods for optimal design of studies for developmental inquiry. *J. Gerontol. Ser. B Psychol. Sci. Soc. Sci.* 73, 113–123. doi: 10.1093/geronb/gbx008

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behav. Res.* 1, 245–276. doi: 10.1207/s15327906mbr0102-10

Chen, Z., and Klahr, D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. *Child Dev.* 70, 1098–1120.

Cheng, P. W., and Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cogn. Psychol.* 17, 391–416. doi: 10.1016/0010-0285(85)90014-3

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: bayesian additive regression trees. *Ann. Appl. Statist.* 4, 266–298. doi: 10.1214/09-AOAS285

Corbett, A. T., and Anderson, J. R. (1995). Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model. User Adapt. Interact.* 4, 253–278. doi: 10.1007/BF01099821

DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educ. Assess.* 12, 23–45. doi: 10.1080/10627190709336946

Díaz-Uriarte, R., and Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinform.* 7:3. doi: 10.1186/1471-2105-7-3

Dietterich, T. (2000). Ensemble methods in machine learning. *Proc. Mult. Classif. Syst.* 1857, 1–15. doi: 10.1007/3-540-45014-9-1

Efron, B. (1978). Regression and ANOVA with zero-one data: measures of residual variation. *J. Am. Statist. Assoc.* 73, 113–121. doi: 10.2307/2286531

Fink, G. A. (2008). *Markov Models for Pattern Recognition*. Berlin: Springer, doi: 10.1007/978-3-540-71770-6

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* 3, 1289–1305.

Freund, Y., and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55, 119–139. doi: 10.1006/jcss.1997.1504

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Statist.* 29, 1189–1232.

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Think. Reason.* 7, 69–89. doi: 10.1080/13546780042000046

Gilula, Z., and Haberman, S. J. (1995). Dispersion of categorical variables and penalty functions: derivation, estimation, and comparability. *J. Am. Statist. Assoc.* 90, 1447–1452. doi: 10.1007/s11336-004-1175-8

Goldhammer, F., Naumann, J., and Keβel, Y. (2013). Assessing individual differences in basic computer skills: psychometric characteristics of an interactive performance measure. *Eur. J. Psychol. Assess.* 29, 263–275. doi: 10.1027/1015-5759/a000153

Goldhammer, F., Naumann, J., Selter, A., Toth, K., Rolke, H., and Klieme, E. (2014). The time on task effect in reading and problem-solving is moderated by task difficulty and skill: insights from a computer-based large-scale assessment. *J. Educ. Psychol.* 106, 608–626. doi: 10.1037/a0034716

Goldstein, B., Polley, E., and Briggs, F. (2011). Random forests for genetic association studies. *Statist. Appl. Genet. Mol. Biol.* 10, 1–34. doi: 10.2202/1544-6115.1691

Goodman, L., and Kruskal, W. (1954). Measures of association for cross classifications. *J. Am. Statist. Assoc.* 49, 732–764. doi: 10.2307/2281536

Greiff, S., Wustenberg, S., and Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem-solving. *Comput. Educ.* 91, 92–105. doi: 10.1016/j.compedu.2015.10.018

Greiff, S., Wüstenberg, S., and Funke, J. (2012). Dynamic problem solving: a new assessment perspective. *Appl. Psychol. Measur.* 36, 189–213. doi: 10.1177/0146621612439620

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182. doi: 10.1162/153244303322753616

Guyon, I., Weston, J., Barnhill, S., and Vapnick, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46, 389–442. doi: 10.1023/A:1012487302797

Haberman, S. J. (1982). Analysis of dispersion of multinomial responses. *J. Am. Statist. Assoc.* 77, 568–580. doi: 10.2307/2287713

Hao, J., Shu, Z., and Davier, A. (2015). Analyzing process data from game/scenario-based tasks: an edit distance approach. *J. Educ. Data Min.* 7, 33–50.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Model Assessment and Selection. The Elements of Statistical Learning*. New York, NY: Springer, 219–259. doi: 10.1007/978-0-387-21606-5-7

He, Q., Borgonovi, F., and Paccagnella, M. (2019). "Using process data to understand adults' problem-solving behaviour in the programme for the international assessment of adult competencies (PIAAC): identifying generalised patterns across multiple tasks with sequence mining," *OECD Education Working Papers* (Paris: OECD Publishing). doi: 10.1787/650918f2-en

He, Q., Glas, C. A. W., Kosinski, M., Stillwell, D. J., and Veldkamp, B. P. (2014). Predicting self-monitoring skills using textual posts on Facebook. *Comput. Hum. Behav.* 33, 69–78. doi: 10.1016/j.chb.2013.12.026

He, Q., Veldkamp, B. P., and de Vries, T. (2012). Screening for posttraumatic stress disorder using verbal features in self-narratives: a text mining approach. *Psychiatr. Res.* 198, 441–447. doi: 10.1016/j.psychres.2012.01.032

He, Q., and von Davier, M. (2015). "Identifying feature sequences from process data in problem-solving items with n-grams," in *Quantitative Psychology Research: Proceedings of the 79th Annual Meeting of the Psychometric Society*, eds A. van der Ark, D. Bolt, S. Chow, J. Douglas, and W. Wang, (New York, NY: Springer), 173–190.

He, Q., and von Davier, M. (2016). "Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment," in *Handbook of Research on Technology Tools For Real-World Skill Development*, eds Y. Rosen, S. Ferrara, and M. Mosharraf, (Hershey, PA: Information Science Reference), 749–776.

He, Q., von Davier, M., and Han, Z. (2018). "Exploring process data in computer-based international large-scale assessments," in *Data Analytics and Psychometrics: Informing Assessment Practices*, eds H. Jiao, R. Lissitz, and A. van Wie, (Charlotte, NC: Information Age Publishing).

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: a conditional inference framework. *J. Comput. Graph. Statist.* 15, 651–674. doi: 10.1198/106186006X133933

Janitza, S., and Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. *PLoS One* 13:e0201904. doi: 10.1371/journal.pone.0201904

Jiang, R., Tang, W., Wu, X., and Fu, W. (2009). A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinform.* 10:S65. doi: 10.1186/1471-2105-10-S1-S65

Jiang, Y., Clarke-Midura, J., Baker, R. S., Paquette, L., and Keller, B. (2018). "How immersive virtual environments foster self-regulated learning," in *Digital Technologies and Instructional Design For Personalized Learning*, ed. R. Zheng, (Hershey, PA: IGI Global.).

Jiang, Y., Paquette, L., Baker, R. S., and Clarke-Midura, J. (2015). "Comparing novice and experienced students in virtual performance assessments," in *Proceedings of the 8th International Conference on Educational Data Mining*, Madrid.

Kim, H., and Loh, W. (2001). Classification trees with unbiased multiway splits. *J. Am. Statist. Assoc.* 96, 589–604. doi: 10.1198/016214501753168271

Kinnebrew, J. S., Mack, D. L., and Biswas, G. (2013). "Mining temporally-interesting learning behavior patterns," in *Proceedings of the 6th International Conference on Educational Data Mining*. Los Altos, CA.

Klieme, E. (2004). "Assessment of cross-curricular problem-solving competencies," in *Comparing Learning Outcomes: International Assessments and Education Policy*, eds J. H. Moskowitz, and M. Stephens, (London: Routledge).

Kohavi, R. (1995). *A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. Los Altos, CA: Morgan Kaufmann.

Kohavi, R., and John, G. (1997). Wrappers for feature selection. *Artif. Intelligence* 97, 273–324. doi: 10.1016/S0004-3702(97)00043-X

Lazonder, A. W., and Rouet, J. F. (2008). Information problem-solving instruction: some cognitive and metacognitive issues. *Comput. Hum. Behav.* 24, 753–765. doi: 10.1016/j.chb.2007.01.025

Lee, Y. H., and Haberman, S. J. (2016). Investigating test-taking behaviors using timing and process data. *Int. J. Test.* 16, 240–267. doi: 10.1080/15305058.2015.1085385

Liao, D., He, Q., and Jiao, H. (2019). Mapping background variables with sequential patterns in problem-solving environments: an investigation of U.S. adults' employment status in PIAAC. *Front. Psychol.* 10:646. doi: 10.3389/fpsyg.2019.00646

Liaw, A., and Wiener, M. (2002). Classification and regression by random forest. *R News* 2, 18–22.

Light, R., and Margolin, B. (1971). An analysis of variance for categorical data. *J. Am. Statist. Assoc.* 66, 534–544. doi: 10.2307/2283520

Lin, Y., and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *J. Am. Statist. Assoc.* 101, 578–590. doi: 10.1198/016214505000001230

Manning, C. D., and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, doi: 10.1.1.121.2604

Martinez, R., Yacef, K., Kay, J., Al-Qaraghuli, A., and Kharrufa, A. (2011). "Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop," in *Proceedings of the 4th International Conference on Educational Data Mining*, Seattle, WA.

Mayer, R. E. (1994). "Problem-solving, teaching and testing," in *The International Encyclopedia of Education*, eds T. Husen, and T. N. Postlethwaite, (Oxford: Pergamon Press).

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., and Levy, R. (2012). Design and discovery in educational assessment: evidence-centered design, psychometrics, and educational data mining. *J. Educ. Data Min.* 4, 11–48.

Mitchell, M. (2011). Bias of the random forest out-of-bag (OOB) error for certain input parameters. *Open J. Statist.* 1, 205–211. doi: 10.4236/ojs.2011.13024

Nicodemus, K., and Malley, J. (2009). Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics* 25, 1884–1890. doi: 10.1093/bioinformatics/btp331

Nigam, K., McCallum, A. K., Thurn, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Mach. Learn.* 39, 103–134. doi: 10.1023/A:1007692713085

Oakes, M., Gaizauskas, R., Fowkes, H., Jonsson, W. A. V., and Beaulieu, M. (2001). "A method based on chi-square test for document classification," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (New York, NY: ACM), 440–441. doi: 10.1145/383952.384080

Oliveira, S., Oehler, F., San-Miguel-Ayanz, J., Camia, A., and Pereira, J. (2012). Modeling spatial patterns of fire occurrence in mediterranean europe using multiple regression and random forest. *Forest Ecol. Manag.* 275, 117–129. doi: 10.1016/j.foreco.2012.03.003

Organisation for Economic Co-operation and Development [OECD], (2014a). *PISA 2012 Results: Creative Problem-Solving: Students' Skills in Tackling Real-Life Problems*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development [OECD], (2014b). *PISA 2012 Technical Report*. PISA. Paris: OECD Publishing.

Peet, R. K. (1974). The measurement of species diversity. *Ann. Rev. Ecol. Syst.* 5, 285–307. doi: 10.1146/annurev.es.05.110174.001441

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* 77, 257–286. doi: 10.1109/5.18626

Ramalingam, D., McCrae, B., and Philpot, R. (2014). "The PISA assessment of problem solving," in *The Nature of Problem Solving*, eds B. Csapó, and J. Funke, (Paris: OECD Publishing), doi: 10.1787/9789264273955-en

Sandri, M., and Zuccolotto, P. (2008). A bias correction algorithm for the Gini variable importance measure in classification trees. *J. Comput. Graph. Statist.* 17, 611–628. doi: 10.1198/106186008X344522

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Shapiro, A. M., and Niederhauser, D. (2004). "Learning from hypertext: research issues and findings," in *Handbook of Research on Educational Communications and Technology*, ed. D. H. Jonassen, (Mahwah, NJ: Lawrence Erlbaum).

Sireci, S., and Zenisky, A. (2006). "Innovative item formats in computer-based testing: In pursuit of improved construct representation," in *Handbook of Test Development*, eds S. Downing, and T. Haladyna, (Mahwah, NJ: Lawrence Erlbaum), doi: 10.4324/9780203874776.ch14

Strobl, C., Boulesteix, A., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measure: illustrations, sources, and a solution. *BMC Bioinform.* 8:25. doi: 10.1186/1471-2105-8-25

Sukkarieh, J. Z., von Davier, M., and Yamamoto, K. (2012). *From Biology to EDUCATION: SCORINg and Clustering Multilingual Text Sequences and Other Sequential Tasks*. Princeton, NJ: Educational Testing Service.

Theil, H. (1970). On the estimation of relationships involving qualitative variables. *Am. J. Sociol.* 76, 103–154. doi: 10.1086/224909

Tschirgi, J. E. (1980). Sensible reasoning: a hypothesis about hypotheses. *Child Dev.* 51, 1–10. doi: 10.2307/1129583

van der Linden, W. (2005). *Linear Models for Optimal Test Design*. New York, NY: Springer, doi: 10.1007/0-387-29054-0

van der Linden, W. J., Klein Entink, R. H., and Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Appl. Psychol. Measur.* 34, 327–347. doi: 10.1177/0146621609349800

Weeks, J. P., von Davier, M., and Yamamoto, K. (2016). Using response time data to inform the coding of omitted responses. *Psychol. Test Assess. Model.* 58, 671–701.

White, A. P., and Liu, W. Z. (1994). Bias in information-based measures in decision tree induction. *Mach. Learn.* 15, 321–329. doi: 10.1007/BF00993349

Winne, P. H., and Baker, R. (2013). The potentials of educational data mining for researching metacognition, motivation and self-regulated learning. *J. Educ. Data Min.* 5, 1–8.

Zaki, M. J. (2001). SPADE: an efficient algorithm for mining frequent sequences. *Mach. Learn.* 42, 31–60. doi: 10.1023/A:1007652502315

Zhang, G., Zhang, C., and Zhang, J. (2010). Out-of-bag estimation of the optimal hyper-parameter in SubBag ensemble method. *Commun. Statist. Simul. Comput.* 39, 1877–1892. doi: 10.1080/03610918.2010.521277

Zhou, M., Xu, Y., Nesbit, J. C., and Winne, P. H. (2010). "Sequential pattern analysis of learning logs: methodology and applications," in *Handbook of Educational Data Mining*, eds C. Romero, S. Ventura, M. Pechenizkiy, and S. J. D. Baker, (Cogent OA: Taylor & Francis), 107–121. doi: 10.1201/b10274-14

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: historical background, methodological developments, and future prospects. *Am. Educ. Res. J.* 45, 166–183. doi: 10.3102/0002831207312909

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership