Research Topics

Validity, reliability and efficiency of comparative judgement to assess student work

Edited by
Sven De Maeyer, Tine Van Daal, Renske Bouwer
Marije Lesterhuis and Eva Hartell

Published in
Frontiers in Education

ISSN 1664-8714
ISBN 978-2-83251-143-5
DOI 10.3389/978-2-83251-143-5

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: frontiersin.org/about/contact

Validity, reliability and efficiency of comparative judgement to assess student work

Topic editors
Sven De Maeyer — University of Antwerp, Belgium
Tine Van Daal — University of Antwerp, Belgium
Renske Bouwer — Utrecht University, Netherlands
Marije Lesterhuis — Spaarne Gasthuis, Netherlands
Eva Hartell — Royal Institute of Technology, Sweden

Citation
De Maeyer, S., Van Daal, T., Bouwer, R., Lesterhuis, M., Hartell, E., eds. (2023). *Validity, reliability and efficiency of comparative judgement to assess student work*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83251-143-5

Table of contents

frontiers | Frontiers in Education

TYPE Editorial
PUBLISHED 12 December 2022
DOI 10.3389/feduc.2022.1100095

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Gavin T. L. Brown,
The University of Auckland,
New Zealand

*CORRESPONDENCE
Tine van Daal
tine.vandaal@uantwerpen.be

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement,
a section of the journal
Frontiers in Education

RECEIVED 16 November 2022
ACCEPTED 28 November 2022
PUBLISHED 12 December 2022

CITATION
van Daal T, Lesterhuis M, De Maeyer S and Bouwer R (2022) Editorial: Validity, reliability and efficiency of comparative judgement to assess student work. *Front. Educ.* 7:1100095. doi: 10.3389/feduc.2022.1100095

Editorial: Validity, reliability and efficiency of comparative judgement to assess student work

Tine van Daal[1]*, Marije Lesterhuis[2], Sven De Maeyer[1] and Renske Bouwer[3]

[1]University of Antwerp, Antwerp, Belgium, [2]University Medical Center Utrecht, Utrecht, Netherlands, [3]Utrecht University, Utrecht, Netherlands

KEYWORDS

comparative judgment (CJ), validity, reliability, efficiency, performance assessment

Editorial on the Research Topic
Validity, reliability and efficiency of comparative judgement to assess student work

Assessing complex skills such as writing, designing, or problem-solving is a challenge. Comparative judgement is considered to be a reliable and valid method for assessing student work (e.g., Lesterhuis et al., 2018; Verhavert et al., 2019). In comparative judgement, students' work is evaluated by pairwise comparison. As assessors only must indicate which piece of work is better, differences in severity are not at play (Pollitt, 2012). Furthermore, each work is compared with several others and evaluated by multiple assessors. Based on these comparisons, the quality of each individual work can be estimated. This quality score reflects, so to speak, the shared consensus of the assessors (Jones et al., 2015; van Daal et al., 2019).

The comparative judgement approach is based on Thurstone's law of comparative judgement (1927), which states that it is possible to discriminate between objects on a single scale through a series of pairwise comparisons (Thurstone, 1927). Even though Thurstone already proposed the possibility of using comparative judgement for assessment in education, it was not until 2004 that Pollitt introduced the method in education in his paper "Let's stop marking exams". His work convincingly explained the merits of comparative assessment in terms of validity and provided the first evidence for a reliable summative assessment. Now, almost two decades later, various comparative judgement tools are available for education, such as Comproved or NoMoreMarking. Moreover, researchers around the world have investigated the quality of the method, where and/or how it can be applied, and how the method can be improved.

In this Research Topic, we aim to provide a state-of-the-art of research on comparative judgement in education. We bring together current insights on the validity, reliability and efficiency of the method. In their contributions to this Research Topic, the authors present recent empirical research, each with their own approach, perspective and research focus. In this way, this Research Topic offers the foundation for future research into comparative judgement.

How valid is comparative judgement?

Up to now, only a limited number of studies dig into the validity of comparative judgement (Whitehouse, 2012; Lesterhuis et al., 2018; van Daal et al., 2019) while this is crucial in light of the use of the scores resulting from comparative judgement (Messick, 1989). More studies into the validity of comparative judgement are highly needed to explore the validity of comparative judgement and factors that might affect the validity of the outcomes (Bejar, 2012).

An important line of research in this Research Topic is focused on the validity of comparative judgement to assess students' competences. Buckley et al. conducted a critical review of how ACJ (adaptive comparative judgement) has been used and studied in the field of technology education. They conclude that there is a need for more critical studies on the internal validity, a theoretical framework, and the consideration of falsifiability. Two studies in this Research Topic add to our knowledge base regarding construct validity, concurrent validity, convergent validity and predictive validity. Mentzer et al. first conducted a content analysis of students' work that was ranked high and low based on a peer assessment making use of CJ, concluding that there is evidence for construct validity. Then they examine the relation between scores obtained through peer assessment, instructors' assessment and students' final grades, concluding that students' peer assessment is an indicator of their final grades (predictive validity) but not an indicator of instructor scores (concurrent validity). Landrieu et al. investigated the extent to which comparative judgement scores converge with absolute analytic and holistic scoring methods. Results show that even though scores generated by the three methods highly correlate, there is substantial variation between methods in the information it gives to researchers and practitioners. This implies that one should consider the goal of an assessment when choosing one of the scoring methods. The authors conclude with an outline on the advantages and disadvantages of each of these methods.

In this Research Topic, there are two studies that investigated specifically the construct validity of comparative judgement. Chambers and Cunningham questioned whether assessors are affected by construct-irrelevant aspects of text quality when comparing texts in an experimental design. They conclude that judgements are influenced by handwriting and the presence of missing responses, showing that some biases might be at play when assessors compare texts. Lesterhuis et al. investigated whether assessors differ in how they evaluate students' work using comparative judgement. More particularly, the authors examined to what extent we can distinguish between different types of assessors based on the aspects they take into account when comparing argumentative texts. Results show that assessors are comparable considering the aspects they evaluate during the process of comparative judgement, but that they are different in the weight they give to some aspects over others. This implies that for valid comparative judgement scores, it is warranted to include multiple assessors.

How reliable is comparative judgement?

Reliability is another important indicator of the quality of an assessment. Most of the early studies on comparative judgement focused on reliability, as it is especially high reliability in which comparative judgement stands out compared to other methods (Pollitt, 2012). In comparative judgement, the scale separation reliability (*SSR*) is used as indicator for reliability (Verhavert et al., 2018). Crompvoets et al., however, questioned this coefficient. They investigated the bias and stability of the *SSR* in relation to the number of comparisons per assessed work based on a simulation study. They conclude that the *SSR* can still be used as an indication of the reliability, even when the variance of the items is overestimated. However, they also recommend to obtain a sufficient number of comparisons per student work (i.e., 41 comparisons per item) to prevent an overestimation of the reliability by the SSR.

How efficient is comparative judgement? New applications, approaches and algorithms

As reliability and efficiency always seem to be a trade-off, it is not surprising that this Research Topic comprises a number of studies on ways to increase efficiency without compromising reliability and validity. Humphry and Bredemeyer show how different sets of works can be efficiently linked using a core set. Verhavert et al. also examined how new student works can be placed in an efficient and reliable manner on a previously calibrated refrence set. They conclude that this alternative application of comparative judgement does not hamper the reliability of scores. Seery et al. outline how CJ can be used as a vehicle to set nation-wide standards and unravel teacher constructs of quality at the same time. Benton describes a simplified pairs approach to increase efficiency in the context of equating standards in high-stakes contexts. His simulation study underpinned its superior accuracy to current approaches. De Vrindt et al. investigated whether and how *text mining* can help to make a CJ assessment of textual products more efficient by taking into account information gained through the text mining in the selection of pairs for CJ. They show that the use of this technique increases efficiency while reducing inflation of the reliability estimate used in CJ. Leech et al. approached CJ as a method for equating the standards set in high-stakes testing contexts, to assure that marks are comparable over the years. They investigate the link between the number and length of tasks and the difficulty of the comparison. They compared the outcomes with the outcome of another method—namely traditional equating—and ask assessors about the judgement processes. They conclude that judges used similar processes in CJ within a topic, but over topics there were differences in how judges come to a decision, making the authors discuss the ability of CJ to maintain an audit of how decisions are made.

Implications for educational practice and future research

This Research Topic shows that applications of comparative judgement widely differs in practice. First, different types of competences and student work are assessed (writing, chemistry, mathematics) demonstrating that the application of comparative judgement is not restricted to a single educational domain. Also, the contributions in this issue show that comparative judgement can be used for different purposes such as peer assessment, instructor assessment, standard setting, and equating. Finally, this Research Topic also demonstrates that the methodology used in research on comparative judgement ranges from qualitative research on assessors' judgement processes, over experimental research to simulation studies. As such, by studying the merits and disadvantages of comparative judgement, the conditions and contexts of comparative judgment have become an interdisciplinary field of research in itself, as demonstrated in this Research Topic.

Author contributions

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

Conflict of interest

ML and SD founded, next to their academic position, Comproved. Comproved makes it possible for teachers to use comparative judgement in their classroom. They have, however, no financial returns from this company.

The remaining 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bejar, I. I. (2012). Rater cognition: implications for validity. *Educ. Meas. Issues Pract.* 31, 2–9. doi: 10.1111/j.1745-3992.2012.00238.x

Jones, I., Swan, M., and Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *Int. J. Sci. Math. Educ.* 13, 151–177. doi: 10.1007/s10763-013-9497-6

Lesterhuis, M., Daal, T., van, Gasse, R. V., Coertjens, L., Donche, V., and Maeyer, S. D. (2018). When teachers compare argumentative texts: decisions informed by multiple complex aspects of text quality. *L1 Educ. Stud. Lang. Literat.* 18, 1–22. doi: 10.17239/L1ESLL-2018.18.01.02

Messick, S. (1989). Meaning and values in test validation: the science and ethics of assessment. *Educ. Res.* 18, 5–11. doi: 10.3102/0013189X018002005

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assess. Educ. Principl. Pol. Pract.* 19, 281–300. doi: 10.1080/0969594X.2012.665354

Thurstone, L. L. (1927). A law of comparative judgment. *Psychol. Rev.* 34, 273–286. doi: 10.1037/h0070288

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., and De Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assess. Educ. Princip. Pol. Pract.* 26, 59–74. doi: 10.1080/0969594X.2016.1253542

Verhavert, S., Bouwer, R., Donche, V., and De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assess. Educ. Principl. Policy Pract.* 26, 541–562. doi: 10.1080/0969594X.2019.1602027

Verhavert, S., De Maeyer, S., Donche, V., and Coertjens, L. (2018). Scale separation reliability: what does it mean in the context of comparative judgment? *Appl. Psychol. Meas.* 42, 428–445. doi: 10.1177/0146621617748321

Whitehouse, C. (2012). *Testing the Validity of Judgements About Geography Essays Using the Adaptive Comparative Judgement Method.* Manchester: AQA Centre for Education Research and Policy.

frontiers
in Education

ORIGINAL RESEARCH
published: 16 December 2021
doi: 10.3389/feduc.2021.772832

Examining the Validity of Adaptive Comparative Judgment for Peer Evaluation in a Design Thinking Course

Nathan Mentzer[1]*, *Wonki Lee*[2] *and Scott Ronald Bartholomew*[3]

[1]*Purdue Polytechnic Institute, Purdue University, West Lafayette, IN, United States,* [2]*College of Education, Curriculum and Instruction, Purdue University, West Lafayette, IN, United States,* [3]*School of Technology, Brigham Young University, Provo, UT, United States*

Adaptive comparative judgment (ACJ) is a holistic judgment approach used to evaluate the quality of something (e.g., student work) in which individuals are presented with pairs of work and select the better item from each pair. This approach has demonstrated high levels of reliability with less bias than other approaches, hence providing accurate values in summative and formative assessment in educational settings. Though ACJ itself has demonstrated significantly high reliability levels, relatively few studies have investigated the validity of peer-evaluated ACJ in the context of design thinking. This study explored peer-evaluation, facilitated through ACJ, in terms of construct validity and criterion validity (concurrent validity and predictive validity) in the context of a design thinking course. Using ACJ, undergraduate students (*n* = 597) who took a design thinking course during Spring 2019 were invited to evaluate design point-of-view (POV) statements written by their peers. As a result of this ACJ exercise, each POV statement attained a specific parameter value, which reflects the quality of POV statements. In order to examine the construct validity, researchers conducted a content analysis, comparing the contents of the 10 POV statements with highest scores (parameter values) and the 10 POV statements with the lowest scores (parameter values)—as derived from the ACJ session. For the criterion validity, we studied the relationship between peer-evaluated ACJ and grader's rubric-based grading. To study the concurrent validity, we investigated the correlation between peer-evaluated ACJ parameter values and grades assigned by course instructors for the same POV writing task. Then, predictive validity was studied by exploring if peer-evaluated ACJ of POV statements were predictive of students' grades on the final project. Results showed that the contents of the statements with the highest parameter values were of better quality compared to the statements with the lowest parameter values. Therefore, peer-evaluated ACJ showed construct validity. Also, though peer-evaluated ACJ did not show concurrent validity, it did show moderate predictive validity.

Keywords: adaptive comparative judgement, comparative judgement, design education, validity and reliability, technology and engineering education

OPEN ACCESS

Edited by:
Tine Van Daal,
University of Antwerp, Belgium

Reviewed by:
Jessica To,
Nanyang Technological University, Singapore
Rosemary Hipkins,
New Zealand Council for Educational Research, New Zealand

****Correspondence:***
Nathan Mentzer
nmentzer@purdue.edu

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *08 September 2021*
Accepted: *09 November 2021*
Published: *16 December 2021*

Citation:
Mentzer N, Lee W and Bartholomew SR (2021) Examining the Validity of Adaptive Comparative Judgment for Peer Evaluation in a Design Thinking Course. Front. Educ. 6:772832. doi: 10.3389/feduc.2021.772832

INTRODUCTION

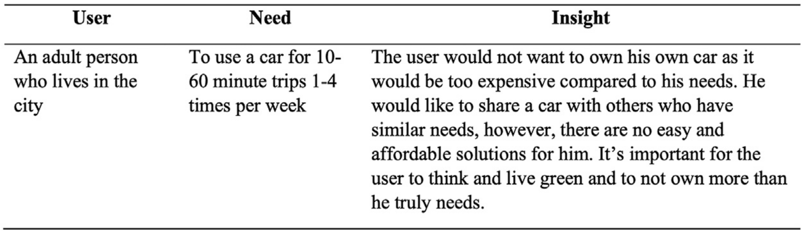
Design is believed to be the core of technology and engineering, which promotes experiential learning towards the development of a robust understanding (Dym et al., 2005; Atman et al., 2008). Design situates learning in real life contexts, involving ambiguity and multiple potentially viable solutions (Lammi and Becker, 2013), and thus promotes the development of students to adapt rapidly to diverse, complicated, and changing requirements (Dym et al., 2005; Lammi and Becker, 2013). Generally, design thinking in the context of technology and engineering settings follows five stages (Erickson et al., 2005; Lindberg et al., 2010): Empathy, define, ideate, prototype, and test. In the stage of empathy, students learn about the users for whom they are designing. Then, they redefine and articulate their specific design problem based on the findings from the empathy stage. Later, students brainstorm creative solutions, build prototypes of ideas, and test prototypes with the original/possible user group to assess their ideas. In the design thinking process, defining the problem is a critical step to capturing what the students are attempting to accomplish through the design. The Point-Of-View (POV) statement (**Figure 1**), which includes three parts (user, need, insight), is one element of problem definition; this artifact often arises during the define stage and serves as a guideline during the entire design process (Sohaib et al., 2019).

In the context of the design thinking course in which this research took place, students worked in groups to write a POV statement to address one or more problem(s) their potential user(s) may confront, by combining user, needs, and insights into a 1-2 sentence statement. Students were instructed that a good problem statement is human-centered, reflecting specific users' insights, broad enough for creative freedom but not too narrowly focused to explore creative ideas, and narrow enough to make it manageable and feasible within a given timeframe (Rikke Friis and Teo Yu, 2020). Hence, a good POV statement is considered a "meaningful and actionable" problem statement (Rikke Friis and Teo Yu, 2020), which guides people to foreground insights about the emotion and experiences of possible user groups (Karjalainen, 2016). It is a crucial step which defines the right challenge to situate the ideation process in a goal-oriented manner (Woolery, 2019) and inspires a team to generate multiple quality solutions (Kernbach and Nabergoj, 2018). Further, effective POV statements facilitate the ideation process by helping an individual to better communicate one's vision to team members or other stakeholders (Karjalainen, 2016).

To encourage students to write well-defined and focused POV statements, design thinking instructors have highlighted the importance of teaching detailed, explicit criteria of good POV statements based on a specific grading rubric (Gettens et al., 2015; Riofrío et al., 2015; Gettens and Spotts, 2018; Haolin et al., 2019). Though competent use of scoring rubrics is believed to ensure reliability and validity of performance assessments, there are inherent difficulties in carrying out rubric-based assessments on summative assignments (Jonsson and Svingby, 2007). Further, this assessment becomes especially difficult in the context of collaborative, project-based design thinking assignments which demand a high level of creativity (Mahboub et al., 2004), especially in terms of organizing the content and structure of the rubric (Chapman and Inman, 2009). Bartholomew et al. have also noted that traditional teacher-centric assessment models (e.g., rubrics) are not always effective at facilitating students' learning in a meaningful way (Bartholomew et al., 2020a) and other studies have raised questions about the reliability and validity of the rubric-based assessment, such as subjectivity bias of the graders (Hoge and Butcher, 1984), one's leniency or severity (Lunz and Stahl, 1990; Lunz et al., 1990; Spooren, 2010), and halo effect due to the broader knowledge of some students (Wilson and Wright, 1993).

In contrast to rubrics, Adaptive comparative judgement (ACJ) has been implemented as an efficient and statistically sound measure to assess the relative quality of each student's work (Bartholomew et al., 2019; Bartholomew et al., 2020a). In ACJ, an individual compares and evaluates pairs of items (e.g., the POV statements) and chooses the better of the two; this process is repeated—with different pairings of items—until a rank order of all items is created (Thurstone, 1927). The pairwise comparison process is iterative and multiple judges can make comparative decisions on multiple sets of work (Thurstone, 1927), with the final ordering of items—from strongest to weakest—calculated using multifaceted Rasch modeling (Rasch, 1980). In addition to a ranking, the judged quality of the items results in the creation of parameter values—which specify both the rank and the magnitude of differences between items—based on the outcome of the judgments (Pollitt, 2012b). Thus, the ACJ approach differs fundamentally from a traditional rubric-based approach in that it allows summative assessment without subjective point assigning (Pollitt, 2012b; Bartholomew and Jones, 2021).

For ACJ, there is no predetermined specific criteria like rubric-based assessments. Rather, in ACJ, holistic statement, or basis for



User	Need	Insight
An adult person who lives in the city	To use a car for 10-60 minute trips 1-4 times per week	The user would not want to own his own car as it would be too expensive compared to his needs. He would like to share a car with others who have similar needs, however, there are no easy and affordable solutions for him. It's important for the user to think and live green and to not own more than he truly needs.

FIGURE 1 | An example of a Point of View (POV) from course reading (Rikke Friis and Teo Yu, 2020).

judgment, is used. This provides the rationale for judges' decisions and is considered a critical theoretical underpinning for reliability and validity (Van Daal et al., 2019). To achieve a level of consensus in ACJ, professionally trained judges' with collective expertise are often considered ideal; however, studies have also demonstrated that students—with less preparation and/or expertise—can also be proficient judges with levels of reliability and validity similar to professionals (Jones and Alcock, 2014). For examples, studies investigating concurrent validity of peer-evaluated ACJ showed that the results generated by peer-evaluated ACJ had a high correlation with the results of experts (e.g., professionally trained instructors, graders) (Jones and Alcock, 2014; Bartholomew et al., 2020a). Jones and Alcock (Jones and Alcock, 2014) conducted peer-evaluated ACJ in the field of mathematics, to see the conceptual understanding of multivariable calculus. The results indicated mean peer and mean expert scores of ACJ had high correlation ($r = 0.77$), and also had significant correlation with summative assessments. Similarly, Bartholomew and others (Bartholomew et al., 2020a) compared the results of professional, experienced instructors' ACJ with student-evaluated ACJ results. Though peer-evaluated ACJ showed non-normality, results suggested strong correlation between peer-evaluated ACJ and instructor-evaluated ACJ.

The present study aims to investigate whether peer-evaluated ACJ can yield sound validity in design thinking. More specifically, the validity of ACJ was studied from two perspectives: construct validity and criterion validity (as investigated through both concurrent and predictive validity). The construct validity was studied based on the holistic nature of ACJ. Three researchers with professional backgrounds evaluated POV statements, studying whether the results of ACJ (parameter values) appropriately reflected general criteria of good POV statement. Following the construct validity, criterion validity was studied. First, researchers investigated concurrent validity of peer-evaluated ACJ by studying the relationships of peer-evaluated ACJ and instructors' rubric-based grading. Second, the researchers studied the predictive validity of peer-evaluated ACJ by studying the relationships of peer-evaluated ACJ and students' final grades. By doing so, we explored the validity of implementing peer-evaluated ACJ in design thinking context.

LITERATURE REVIEW

In this section, we first will start by introducing the concept of a POV statement and the importance of a good POV statement in a design thinking context. Then, two assessments implemented to evaluate POV statements will be presented: rubric-based grading and ACJ. To explore the potential of ACJ as an effective and efficient alternative to rubric-based grading widely implemented in design thinking context, we share a brief review of existing literature on the reliability and validity prior to making our contribution to the knowledge base through this research.

Point-Of-View Statements

The problem definition stage of design thinking explores the problem space and creates a meaningful and actionable problem statement (Rikke Friis and Teo Yu, 2020). Dam and Siang asserted that a good POV statement has three major traits (Dam and Siang, 2018). First, the POV needs to be human-oriented. This means the problem statement students write should focus on the specific users, from whom they learn the needs and insights through the empathy stage. Also, a human-centered POV statement is required to be about the people who are stakeholders in the design problem rather than the technology, monetary return, and/or product improvement. Second, the problem statement should be broad enough for creative freedom meaning the problem statement should be devoid of a specific method or solution. When the statement is framed around a narrowly defined solution, or with a possible solution in mind, it restricts the creativity of the ideation process (Wedell-Wedellsborg, 2017). The final trait of a strong problem statement is that it should be narrow enough to make it viable with the available resources. The third trait complements the second trait, which suggests that the POV statements should possess appropriate parameters for the scope of the problem, avoiding extreme narrowness or ambiguity. A good POV statement, equipped with all three traits, can contribute to delivering attention, providing sound framework for the problem, motivating students working on the problem, and providing informational guidelines (Sohaib et al., 2019).

Assessment of Point-Of-View Statements With Rubrics

One trend among assessments in higher education is a shift from traditional knowledge-based tests towards assessment to support learning (Dochy et al., 2006). In order to capture students' higher-order thinking, a credible, trustworthy assessment, which is both valid and reliable, is needed. The historic development of a rubric as a scoring tool for the assessment of students' authentic and complex work, including what counts (e.g., user, needs, insights are what count in POV statements) and for how much, has traditionally centered on 1) articulating the expectations of quality for each task and 2) describing the gradation of quality (e.g., excellent to poor, proficient to novice) for each element (Chapman and Inman, 2009; Reddy and Andrade, 2010). Three factors are included in a rubric: evaluation criteria, quality definitions, and a scoring strategy. The analytic rubric used in the Design thinking course to grade POV statements is included below (**Table 1**). The rubric-based evaluation of competency is made through analytical reflections by graders, in which the representation of the ability is scored on a set of established categories of criteria (Coenen et al., 2018).

Adaptive Comparative Judgment

Adaptive comparative judgment (ACJ) is an evaluation approach accomplished through multiple comparisons. In 1927, Thurstone presented the "Law of Comparative Judgment" (Thurstone, 1927) as an alternative to the existing measurement scales, aimed at increasing reliability. Thurstone specifically argued that making decisions using holistic comparative judgments can increase reliability compared to decisions made from predetermined rubric criteria (Thurstone, 1927). Years later, based on

TABLE 1 | Grading rubrics of POV statements from the design thinking course.

Evaluation criteria	Proficient	Adequate	Novice	Criterion score
Detail for USER and NEEDS	(6 points) Student work includes adjectives and details to describe the users and their needs. 1 USER and 1 NEED are identified. USERS and NEEDS are clear and concise, actionable, and provide a solid framework for a problem	(3 points) Fewer than the required number of USERS and NEEDS have been generated. USERS and NEEDS are too vague to be useful	(0 points) None	6
INSIGHT	(4 points) Student work shows evidence of considering multiple insights based on the USER and NEEDS. INSIGHTS are surprising and inspirational	(1 point) Evidence for only single INSIGHT was shown. INSIGHT is not based on the USERS or NEEDS; they are uninspiring or obvious	(0 points) None	4
POV	(5 points) Students generated 1 POV statement stemming from the USERS and NEEDS generated. The statement is synthesized, clear, and actionable	(2.5 points) USER, NEED, and INSIGHT are not aligned with each other or the problem. The POV is too vague to be useful, it is unclear, and/ or not actionable	(0 points) None	5

Thurstone's law of comparative judgement, Pollitt outlined the potential for ACJ, seeking the possibility of implementing the comparative judgment approach in marking a wide range of educational assessments (Pollitt, 2012b), with statistically sound measurements in terms of accuracy and consistency (Bartholomew and Jones, 2021). The adaptive attribute of ACJ is based on an algorithm embedded within the approach which pairs similarly ranked items as the judge makes progress in the comparative judgement process—an approach aimed at expediting the process of achieving an acceptable level of reliability (Kimbell, 2008; Bartholomew et al., 2019).

We choose to use a software titled RMCompare to facilitate adaptive comparative judgment enabling students to make a series of judgments with an outcome consisting of several helpful data, including: a rank order of the items judged, parameter values (statistical values representing the relative quality of each item), judgment time of each comparison, a misfit statistic of judges and items (showing consistency, or lack thereof, among judgments), and judge-provided rationale for the comparative decisions (Pollitt, 2012b). Previous research has shown that utilizing these data can provide educators with a host of possibilities including insight into students' judgment criteria, consensus, and their processing/ understanding of the given task. In a design thinking process scenario specifically, ACJ—though originally designed for expert assessment—has demonstrated through educational research efforts to be a helpful measure for students who participate in the task because it promotes learning and engagement (Seery et al., 2012; Bartholomew et al., 2019). Specifically, Bartholomew et al. noted that ACJ can efficiently facilitate learning among students studying design and innovation by including students as judges (Bartholomew et al., 2020a).

Validity of Adaptive Comparative Judgment

Construct Validity of Adaptive Comparative Judgment: Holistic Approach

The traditional concept of validity was established by Kelley (Kelley, 1927), who claimed that validity is the extent to which a test measures what it is supposed to measure. Construct validity pertains to "the degree to which the measure of a content sufficiently measures the intended concept" (O'Leary-Kelly and Vokurka, 1998, *p*. 387). The validity estimate has to be considered in the context of its use, and needs evidence of the relevance and the utility of the score inferences and actions (Messick, 1994). In other words, researchers need to take into account the context, with adequate construct validity evidence, to support the inferences made from a measure (Hubley and Zumbo, 2011).

Since ACJ requires holistic assessment, researchers examining the validity of comparative judgement have highlighted the importance of an agreed upon set of criteria (Pollitt, 2012a) and shared consensus across judges (Pollitt, 2012a; Jones et al., 2015; Van Daal et al., 2019). In terms of an agreed upon criteria for judgment, in some instances, rather than following a predetermined specific criterion for the assessment, judges in ACJ have followed a general description regarding the assessment. For instance, Pollitt (Pollitt, 2012a) used the "Importance Statements" published on England's National Curriculum to assess design thinking portfolios:

> In design and technology pupils combine practical and technological skills with creative thinking to design and make products and systems that meet human needs. They learn to use current technologies and consider the impact of future technological developments. They learn to think creatively and intervene to improve the quality of life, solving problems as individuals and members of a team.
>
> Working in stimulating contexts that provide a range of opportunities and draw on the local ethos, community and wider world, pupils identify needs and opportunities. They respond with ideas, products and systems, challenging expectations where appropriate. They combine practical and intellectual skills with an understanding of aesthetic, technical, cultural, health, social, emotional, economic, industrial, and environmental issues. As they do so, they evaluate present and past design and technology, and its uses and effects. Through design and technology pupils develop confidence in using practical skills and become

> discriminating users of products. They apply their creative thinking and learn to innovate. (QCDA., 1999).

The shared consensus among judges, facilitated through the ACJ process, underpins the validity of ACJ, because each artifact is systematically evaluated in various pairings across multiple judges. Through the process of judgement, a shared conceptualization of quality and collective expertise of judges is then reflected in the final rank order (Van Daal et al., 2019). Though the majority of studies initially limited the judges to trained graders/instructors, recent work has explored students' (or other untrained judges') competence as judges in ACJ (Rowsome et al., 2013; Jones and Alcock, 2014; Palisse et al., 2021). Findings suggest that, in many cases, students—and even out-of-class-professionals (e.g., practicing engineers; see Strimel et al., 2021) can reach similar consensus to that reached by trained judges or classroom teachers suggesting a shared quality consensus across different judge groups.

Considering the curriculum, goals, and educational setting of design thinking, our research team postulated that when implementing ACJ to assess POV statements of the students in the design thinking course, the high score of parameter values should reasonably be interpreted as one's ability to write a good POV statement, while a low score of parameter values can be understood as one's low ability, or lack of ability, to write a good POV statement.

Validity of Adaptive Comparative Judgment: Criterion Validity

In classical views of validity, criterion validity concerns "the correlation with a measure and a standard regarded as a representative of the construct under consideration" (Clemens et al., 2018). If the measure shows a correlation with an assessment in the same time frame, it is termed concurrent validity. If the measure shows a correlation with a future assessment, it is termed predictive validity. The criterion validity evidence is related to how accurately one measure predicts the outcome of another criterion measure. Criterion validity is useful for predicting performance of an individual in different context (e.g., past, present, future) (Borrego et al., 2009).

Although the unique, holistic characteristics of ACJ provides meaningful insights, concurrent validity of ACJ also has been studied with great importance (Jones and Alcock, 2014; Jones et al., 2015; Bisson et al., 2016). There has been several efforts to establish criterion validity of ACJ, which mostly concentrated on the concurrent validity (Jones and Alcock, 2014; Jones et al., 2015; Bisson et al., 2016). These studies compared the results of ACJ with the results of other validated assessments to investigate the conceptual understanding. Examining the criterion validity is crucial to implement ACJ in various educational contexts as an effective alternative. Considering that ACJ can be rapidly applied to target concepts, it has the potential to effectively and efficiently evaluate various artifacts in a wide range of contexts with high validity and reliability (Bisson et al., 2016).

Informed by previous studies, this study examines the validity of peer-evaluated ACJ in design thinking context. Though it has relatively high and stable reliability, coming from its adaptive nature, empirical evidence regarding ACJ's predictive validity is limited (Seery et al., 2012; Van Daal et al., 2019). Delving into predictive validity is necessary for demonstrating the technical adequacy and practical utility of ACJ (Clemens et al., 2018). Therefore, investigating the validity of ACJ may provide another potentially strong peer assessment measure in design thinking context, where most of the assignments are portfolios, thus hard to operationalize explicit assessment criteria using traditional rubric based approaches (Bartholomew et al., 2020a). Not only may ACJ be a viable assessment tool but, it may also be a valuable learning experience for students who engage in the peer evaluation process (Bartholomew et al., 2020a).

RESEARCH QUESTION

The ACJ-produced rank order and standardized scores (i.e., parameter values) reflect the relative work quality of students' POV statements according to the ACJ judges. Therefore, researchers assumed that POV statements with higher parameter values were better in quality when compared to the POV statements with lower parameter values. The first research question investigated in this study will qualitatively explore how students' shared consensus reflects the general and broad criteria of good POV statement.

RQ 1. What is the construct validity of ACJ? Does peer-reviewed ACJ reflect general criteria of good POV statements?

Taking its effectiveness and efficiency into consideration, studies already explored ACJ's theoretical promise in educational setting as a new approach with acceptable statistical evidence (Jones and Alcock, 2014; Bartholomew et al., 2020a). This study aims to investigate the criterion validity of ACJ. More specifically, concurrent validity and predictive validity of ACJ were examined by comparing the results of ACJ with rubric-based grading.

RQ 2. What is the criterion validity of ACJ? Does peer-reviewed ACJ correlate with existing assessment?

RQ 2-1. What is the concurrent validity of ACJ? Does peer-reviewed ACJ correlate with instructors' rubric-based grading on the same assignment?

RQ 2-2. What is the predictive validity of ACJ? Does peer-reviewed ACJ predict instructors' rubric-based grading on the key final project deliverable?

METHODS

Participants

Study participants were 597 technology students out of 621 students enrolled in a first-year Design Thinking Course at a large Midwestern university in the United States during Spring 2019. These students are subset of entire Polytechnic population (N = 4,480). This research was approved by the university's Institutional Research Board. Sociodemographic information of the participants is provided in **Table 2**.

TABLE 2 | Sociodemographic characteristics of participants.

Socio-demographic variables	Number	Percent
Gender		
Female	147	24.62
Male	446	74.71
Prefer not to answer	4	0.67
Residency		
Foreign	52	8.72
Non-Resident	207	34.67
Resident	334	55.95
Prefer not to answer	4	0.67
Race		
Multiracial	17	2.85
Alaskan Native	1	0.17
Asian	53	8.88
Black/African American	14	2.35
Hispanic/ Latino	41	6.87
Native American	1	0.17
Unknown	8	1.34
White	406	68.01
Prefer not to answer	56	9.38
Rank by credit hour		
Freshman	182	30.49
Sophomore	235	39.36
Junior	124	20.77
Senior	52	8.71
Prefer not to answer	4	0.67

Research Process

Research Design

The research design of this study is graphically depicted by **Figure 2**. First, students wrote the POV statements during the project 3 as a team. Researchers collated and anonymized the total 124 POV statements. Followed by this process, students performed ACJ on their peer's POV statements (Assessment 1, peer-evaluated ACJ). Concurrently, instructors graded the same POV statement using rubrics (Assessment 2, **Table 1**). After project 3, instructors, who worked as graders assigned grades to final deliverables of project 3 (Assessment 3). To study the construct validity, researchers qualitatively analyzed ACJ statements using content analysis. Before analyzing the criterion validity, we analyzed the descriptive statistics of all three assessments. For the concurrent validity, we studied correlation between the peer-evaluated ACJ (Assessment 1) and instructors' grading based on rubric (Assessment 2). Finally, for the predictive validity, we examined if peer-evaluated ACJ (Assessment 1) predicts final deliverables (Assessment 3).

Study Context and Point-Of-View Statement Writing

In the semester-long, three credit design thinking course, 597 students from 14 sections designed and developed solutions to

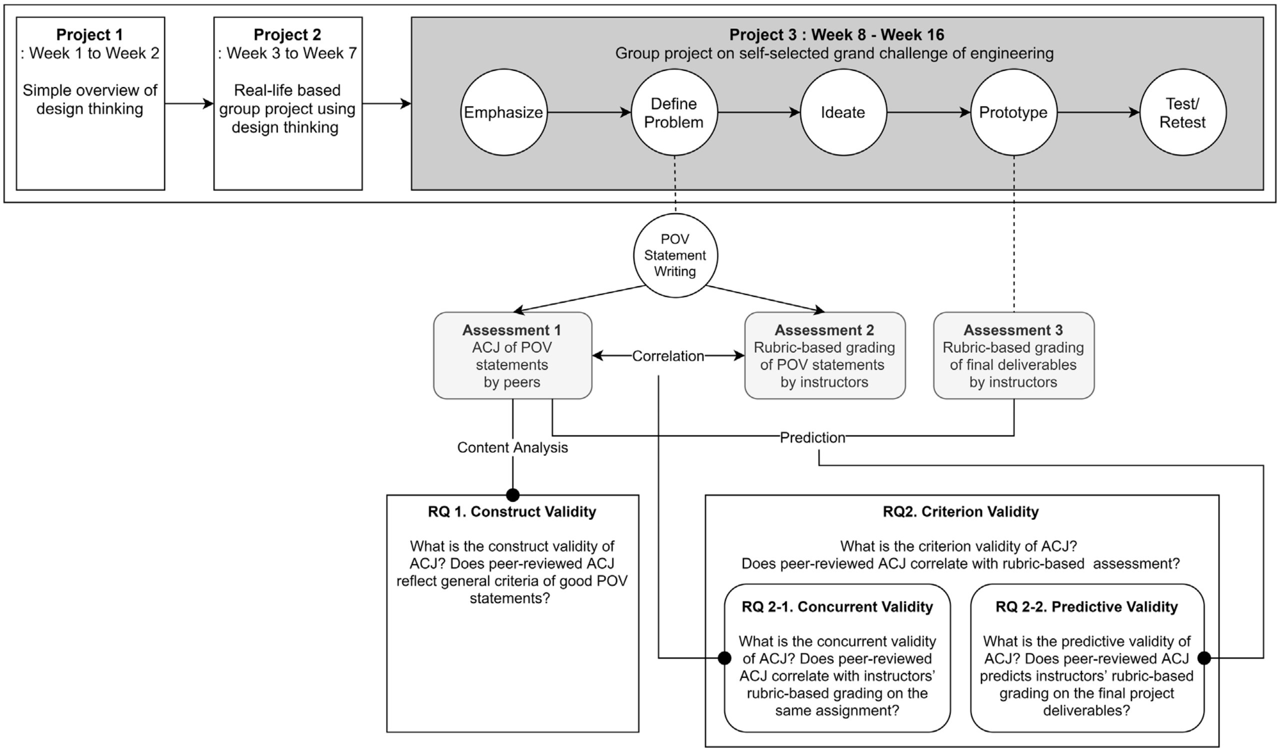

Project 1
: Week 1 to Week 2
Simple overview of design thinking
Project 2
: Week 3 to Week 7
Real-life based group project using design thinking
Project 3 : Week 8 - Week 16
Group project on self-selected grand challenge of engineering
Emphasize
Define Problem
Ideate
Prototype
Test/ Retest
POV Statement Writing
Assessment 1
ACJ of POV statements by peers
Correlation
Assessment 2
Rubric-based grading of POV statements by instructors
Assessment 3
Rubric-based grading of final deliverables by instructors
Content Analysis
Prediction
RQ 1. Construct Validity
What is the construct validity of ACJ? Does peer-reviewed ACJ reflect general criteria of good POV statements?
RQ2. Criterion Validity
What is the criterion validity of ACJ?
Does peer-reviewed ACJ correlate with rubric-based assessment?
RQ 2-1. Concurrent Validity
What is the concurrent validity of ACJ? Does peer-reviewed ACJ correlate with instructors' rubric-based grading on the same assignment?
RQ 2-2. Predictive Validity
What is the predictive validity of ACJ? Does peer-reviewed ACJ predicts instructors' rubric-based grading on the final project deliverables?

FIGURE 2 | Research design of this study.

real problems, voluntarily forming 124 groups in alignment with their current interests or major within each section of the course. During the course, students fostered their own foundational understanding of design thinking by participating in three projects, in which they could create, optimize, and prepare innovative solutions for people. The first project was designed to provide overview and theoretical descriptions with simple hands-on projects about the design thinking process and lasted about a week. The second course project was a more real-life based group project, and took approximately 4 weeks, following the five stages of design thinking: empathize, define the problem, ideate, prototype, and test (retest).

The final project spanned about 8 weeks and engaged students in addressing a problem related to a self-selected grand challenge of engineering (National Academy of Engineering, 2008). In this study, we observed the "define" stage of the third project, when we hypothesized that students would have had enough experience with the design thinking process, including the POV statements, to work comfortably through the designing approach. At this point in class these students had already written four POV statements, two as an individual during the first project, and two as a team during the second project. As a part of the define stage during the third project, the course instructors utilized one 50-min class concentrating on POV creation, highlighting essential components of quality POV statements (user, needs, and insights), structures of POV statements, essential criteria for producing a good POV statement, and importance of writing a good POV statement for this project. During and after this class session, the students wrote a definition of their problem as a team using a provided format for POV statements [User . . . (descriptive)] needs [need . . . (verb)] because [insight. . . (compelling)].

Measures

This study used three types of assessments: peer-evaluated ACJ of POVs (Assessment 1), rubric-based grading of POV(Assessment 2), and rubric-based grading of final deliverables (Assessment 3). First, we compared two types of assessments: Assessment 1 and Assessment 2. For both rubric based and ACJ based assessments, all the POV statements from the 124 teams written at the beginning of the final project were included in the dataset. Then, researchers included the rubric-based grading of final deliverables (Assessment 3) to see if the peer-evaluated ACJ can predict the future achievements.

Assessment 1. Peer-Evaluated ACJ of the POV Statements.

For the peer-evaluated ACJ, the POV statements were collated, anonymized, and uploaded into the ACJ software called *RMCompare* for evaluation. Near the end of the final project, in preparation for presenting their design projects, students were challenged to evaluate the POV statements using the *RMCompare* interface by selecting the POV statement they believed was holistically better between the pairs displayed to them. For the holistic judgment prompt, students were reminded of general qualities of good POV statements (Rikke Friis and Teo Yu, 2020), which were already familiar to them. Students previously used these same criteria (Rikke Friis and Teo Yu, 2020) as class material to learn the notion of POV statement. Each student (550 of 597) compared approximately 8 pairs of POV statements written by their peers. The subsequent ACJ judgments resulted in all 124 POV statements being compared at least 12 times to other increasingly similarly ranked POV statements in line with the adaptive nature of the software. As a result, the rank and parameter value for each POV statement was automatically calculated using the embedded Rasch multifaceted model (see Pollitt, 2012b; Pollitt, 2015 for more details).

Assessment 2. Instructor's Rubric-Based Grading of the POV Statements.

Rubric based grading was performed based on assigned criteria (**Table 1**). Graders are currently working as course instructors of design thinking course, who were pursuing a MS or Ph.D. degree in relevant fields (e.g., engineering, polytechnic, or education) at the time of study. Each grader assessed two sections, in which around 40 students enrolled. As a result, the numerical grading value (total 15 pts) were provided.

Assessment 3. Final Project Deliverables.

Student teams submitted their final prototypes as one of the significant final project deliverables. They plan, implement, and reflect on testing scenarios for their prototypes, and present prototypes for the purpose of receiving feedback from the peers. Instructors (same as Assessment 2) grade the prototypes as a key final deliverable based on assigned criteria (see **Table 3**). As a result, the numerical grading value (total 35 pts) were provided.

Analysis

Construct Validity

Qualitative Content Analysis (QCA)

Content analysis is an analytic method frequently adopted in both quantitative and qualitative research for the systematic reduction of text or video data (Hsieh and Shannon, 2005; Mayring, 2015). Qualitative content analysis, QCA is one of the recognized research methods in the field of education. It is a method for "the subjective interpretation of the content of text data through the systematic classification process of coding and identifying themes or patterns" (Hsieh and Shannon, 2005, *p*. 1278). We used directive (qualitative) content analysis to extend the findings of ACJ, therefore enriching the findings (Potter and Levine-Donnerstein, 1999). The focus of current study was on validating ACJ from analyzing the key concepts of POV statements (e.g., structure, user, needs, and insights). Researchers began the research by identifying the key concepts POV statements. Then, researchers begin coding immediately with the predetermined codes. We articulated four categories based on the discussion: framework (alignment, logic), user, needs, and insights.

Two major approaches are frequently used for the validity and reliability of QCA: Quantitative and qualitative (Mayring, 2015). Quantitative approach measures inter-coder reliability and agreement using the quantitative methods (Messick, 1994). Qualitative approach adopts a consensus process in which multiple coders independently code the data, compare their coding, and discuss and resolve discrepancies when they arise, rather than measuring them (Schreier, 2012; Mayring, 2015). The qualitative validation approach is preferred to the quantitative

TABLE 3 | Rubrics of the final project deliverable.

Criteria	Proficient	Adequate	Novice	Criterion score
Sketches of how it will work provided	(4 points) Sketches illustrating how it works	(2 points) Sketches provided for prototype are provided but are misaligned and/or unclear	(0 points) Sketches entirely lacking	4
Area of concern/ functionality investigated by prototype described	(4 points) Robust description provided for prototype	(2 points) Descriptions are provided but muddled/ unclear	(0 points) Insufficient descriptions provided	4
Picture of prototype included; Description of how prototype was built included	(4 points) Pictures and robust description provided for prototype	(2 points) Some pictures provided; descriptions are provided but muddled/unclear	(0 points) Picture lacking; Insufficient descriptions provided	4
Pictures provided of prototype "in use"; description of relevant test conditions	(4 points) Pictures and robust description provided for prototype	(2 points) Pictures included; descriptions provided for prototype; descriptions are provided but muddled/unclear	(0 points) Pictures lacking; Insufficient descriptions provided	4
Test results provided	(5 points) Test results included; results are primarily quantitative with supplemental qualitative results included	(2.5 points) Test results included but results primarily observational or anecdotal	(0 points) Test results either lacking, or extremely insufficient	5
Most comparable existing product pictured; differences described	(4 points) Pictures included; differences provided	(2 points) Pictures provided; differences provided but are muddled/unclear	(0 points) Pictures lacking; Insufficient differences provided	4
Prototype Functions	(10 points) The group's prototype functions properly	(5 points) The prototype partially function	(0 points) The prototype does not function	10

research because it provides reason with reflexivity, the critical thinking of researchers' own assumptions and perspective (Schreier, 2012). This is particularly important during the negotiation process because coders meet to discuss their own rationale used in coding. In this study context, researchers compared, reviewed, and revisited coding process before reaching consensus on the codes (Hsieh and Shannon, 2005; Forman and Damschroder, 2007; Schreier, 2012).

Sample Selections of Point-Of-View Statements

To provide validation to ACJ data (parameter values), researchers selectively analyzed 20 POV statements out of the 124 POV statements as was done in a previous related study (Bartholomew et al., 2020b). Based on ACJ, we selectively analyzed the 10 POV statements with the highest parameter values and the 10 POV statements with the lowest parameter values to provide contrasting cases. Using the rubrics implemented in the grading system (**Table 1**), researchers analyzed whether the parameter values were aligned with the criteria for a strong POV statement. More specifically, in an effort to explore the construct validity of the ACJ results, we investigated if the 10 POV statements with high parameter values better reflect the required criteria for good POV statements and if the 10 POV statements with low parameter values fail to meet the criteria required of the student groups.

Criterion Validity Analysis

The software program RStudio Version 1.3.959 was used for our criterion validity analysis.

Preliminary Data Analysis

Prior to running the statistical analysis, researchers screened the data for missing values and outliers. Participants with missing data on a variable were excluded from the analysis. For instance, if there was a missing value either in grader's grading in POV statements or final deliverables, the data were not included in the statistical analysis. As a result, 26 participants were removed from data. Values greater than 4 SD from the mean on any measures were considered as outliers and thus removed. The results of ACJ demonstrated a high level of interrater reliability ($r = 0.94$), with none of the judges showing significant misalignment.

Descriptive Statistics

We analyzed the rubric based grading of POV statements (POV Grading), ACJ on the same POV statements (ACJ), and rubric-based grading on the final deliverables (Final Deliverable) (**Table 4**).

Correlation and Regression Analysis

Specifically, both Spearman's ρ and linear regression statistical techniques were employed to test the concurrent validity and predictive validity. We adopted Spearman's ρ because the POV grading was negatively skewed.

TABLE 4 | Descriptive statistics.

	N	Min	Max	Mean	SD	Skewness		Kurtosis	
						Statistic	SE	Statistic	SE
POV Grading	576	0.00	15.00	13.08	3.74	−2.70	0.10	6.62	0.20
ACJ	576	−1.80	1.23	0.01	0.56	−0.53	0.10	0.68	0.20
Final Deliverable	576	16.25	35	20	2.65	1.78	0.10	6.94	0.21

TABLE 5 | POV statements with the highest parameter values.

Rank order	Point-of-view statement	Parameter value
#1	The school of aviation and transportation technology needs to utilize a more accessible, personalized and interactive method for giving safety meetings because currently they lack motivation and differently levels of complexity within the class environment	1.23
#2	People utilizing automobiles and transportation vehicles need a way to reduce the amount of CO_2 emissions of the transportation sector because the transportation sector is the largest emitter of CO_2 as of 2018, which leads to more impacts of global warming	1.17
#3	College aged students need a way to learn about the importance of recycling, by reusing wasted materials in an effective manner because it will reduce the carbon footprint that college campuses leave	0.94
#4	The people of (Name of the City) should be offered an incentive to recycle responsibly, because recycling is being done the wrong way which hurts the environment more than it helps	0.91
#5	College students need technology and social networks as an alternative form of learning about reading mainly in English classes so that students have access to alternative forms of non-discriminatory educational methods	0.90
#6	(Name of the University) students need a way of navigating (Name of the University's) flooded sidewalks without getting their feet soaked in snow or ice because walking into class with cold and wet boots because it is both unsanitary and potentially dangerous, especially in the winter months	0.87
#7	Due to time and accessibility constraints, students on campus need a means to achieve a healthier lifestyle without spending too much extra time and money, because better health is very important to busy and stressed college students	0.87
#8	Junior High students need an interactive method of teaching fundamental ideas of STEM because the current system of teaching lacks the support, motivation, and exposure students need to grow intellectually	0.81
#9	University members need a consistently secure authentication service because hacked accounts can lead to data leakage and theft	0.79
#10	Local business owners need a cheap and efficient way to cool their data lefts, and reuse the energy because the current technology involving air conditioning and water cooling is very expensive and wasteful to the environment	0.76

**Note: Original statements are as written by students.*

TABLE 6 | POV statements with the lowest parameter values.

Rank order	Point-of-view statement	Parameter value
#115	People need to become more educated on the topics of stereotyping and cultural diffusion because ignorance can lead to discrimination	−0.85
#116	People in the (Name of the University) university need assist to find parking spots because currently there is no helpful approach improve the shortage of parking slots	−0.86
#117	People at (Name of the University) University do not have access to cheap, healthy food for an unknown reason	−0.88
#118	Pedestrians need signage to prevent vehicle users in the bike lanes from hitting them because there is a high risk of accidents in that area	−0.97
#119	Anyone involved in scientific or technological labs currently have no access to virtual lab spaces to practice techniques or methods that are otherwise difficult to obtain physically	−0.98
#120	The VR market is growing rapidly since 2012, but it has not yet reached a mature market. We are going to explore challenges Virtual Reality needs to overcome in order to be more adaptable for people, especially for educational purposes. People who are in the education system need a way to incorporate Virtual Reality into teaching and learning because VR provides a new way to share immersive information in an affordable way	−1.00
#121	People who live in urban areas need a sustainable source of foods because it decreases their reliance on imports	−1.13
#122	The Food Industry needs to waste less because the environment is suffering due to excessive usage of natural resources	−1.35
#123	We will implement lights above each parking spots in parking garage, and they will glow either green or red depending on whether it's available or not	−1.69
#124	Infrastructure at (Name of the University) needs to be improved because parts of (Name of the University) are overcrowded	−1.80

**Note: Original statements are as written by students.*

RESULTS

Construct Validity of Peer-Evaluated Adaptive Comparative Judgment

The POV statements with the highest parameter values (**Table 5**) and the lowest parameter values (**Table 6**) are presented based on their rank order and referenced in the following discussion.

Framework of Point-Of-View Statements

Structure and Length

To articulate their user, needs, and insights to solve the current challenges users are facing, the assignment required students to make a POV statement using the sentence structure: [User . . . (descriptive)] needs [need . . . (verb)] because [insight. . . (compelling)] (Rikke Friis and Teo Yu, 2020). Though most of the POV statements with high parameter values followed the basic structures, some of the POV statements with low parameter values deviated from the basic POV statement structure. For instance, the POV #117 and #119 statements omitted insights resulting in their POV statements not leading to an actionable statement. The #120 statement included unnecessary background information prior to the POV statement which may be distracting and hinder the readers' understanding of the POV statement itself. In the #123 statement, a specific solution was presented instead of the POV statement and a problem statement like this, framed with a certain solution in mind, might restrict the creativity of problem-solving (Wedell-Wedellsborg, 2017). Therefore, based on our analysis, the judges perceived that good POV statements should include the required information with all the necessary components (i.e., user, needs, insights) in a concise manner with the necessary details.

In terms of the length, researchers found the POV statements of low parameter values were notably shorter than the POV statements with high parameter values, except for the statement #120. It provides insights to the researchers that the students produced POV statements with lower parameter values are not clearly specifying the user, need and insight. Therefore, short length reflects the lack of thorough description to understand the context in which the POV statements are based on. Also, when we took a more detailed analysis on the statement #120, we found that this statement included introductory sentence as part of their POV statement. The inclusion of introductory sentences can either be interpreted as students' misunderstanding of the structure of POV statement, or lack of writing skills to integrate all the necessary detailed information in the structure of POV statement.

Alignment and Logic

The user, needs, and insights should be aligned and actionable to increase the likelihood of success during the follow-up designing process. Well-aligned POV statements enhance the team's ability to assist the users in meeting their goals and objectives in an efficient and effective way (Wolcott et al., 2021). Compared to the high parameter value statements, our research team agreed that the low parameter value statements typically showed less logically aligned user, needs, and insights. In most of the cases, the less cohesive POV statements came from stating the user and needs in a manner that was too broad, vague, or less clarified. Statement #121, #122, #124 were direct examples of this problem. For instance, the statement #121 fell short of a detailed illustration about why "people who live in urban areas" needed a "sustainable source of foods". Too broad of a user group, like "people live in urban areas", was not cohesively related to the need of "sustainable foods", and this statement did not articulate what were the "sustainable foods". Thus, it appeared difficult to determine whether it was hard to gain sustainable sources of food in urban areas, or whether the struggles were due to the socio-economic status of the residents in urban districts that more sustainable sources of food were needed. Moreover, the insights did not clarify the range and definition of "imports", and why it was important and/or positive to decrease the reliance on imports.

POV statements lacking alignment between the user, need and insight were not logical and/or easy to follow. These kinds of statements appeared unfounded or unsupported. For instance, statement #117, #119, #120, #121, and #122 could face rebuttal because the user group was not well aligned with the needs. As an example, the statement #122 insisted that the "Food industry" "waste less", to prevent "excessive usage of natural resources". Not only were the contents of this statement not written in the way POV statements required, but it also lacked a logical explanation of why the food industry needed to waste less, while there could be many possible factors/ subjects excessively wasting natural resources. Overall, not including the components of a POV statement (user, need and insight) or including them in ways that are not well aligned yield POV statements that are marginally actionable and vague. Additionally, the lower quality POV statements often framed the users' needs as oriented towards a specific solution rather than focusing on the problem at hand.

Components of Point-Of-View Statements

User

Although these were broad in some senses, the user defined in both the POV statements with high parameter values and low parameter values were narrowed down with descriptive explanations, though the degree of specification differed from statement to statement. Specifically, some of the POV statements with low parameter values revealed limitations when defining users. For instance, the statement #115 defined "People" as a user group but did not narrow down the user and not provide any illustrated details about the user group they are targeting. The user group of the statement #118 was "pedestrians", which was not any different from "people", failing to narrow it down enough. The statement #123 did not designate any user group, therefore making the targeted user group remain unspecified. By failing to define user groups from the specific user's perspective in the problem-solving, these teams fell short of solutions with quantity and higher quality.

Needs

The needs are something essential or important, and are required for targeted users (Interaction Design Foundation, 2020). Though it still could have been improved, compared to the low parameter value statements, most of the high parameter value statements incorporated adjectives and details specific to the user group. For instance, the statement #1 and #2 proposed the needs pertinent to the user group. The statement #1 proposed a need for an "accessible, personalized and interactive" method for safety meetings. When limited to the user and needs, this statement did not seem to provide sufficient information due to the vague depiction of the user group. However, considering their insights illustrated the current situation of the statement #1 user group, it seemed to reflect the current needs the user group was confronting. The statement #2 also showed needs of "reducing the CO2 emissions" relevant to the user group utilizing the automobiles and transportation vehicles. Also, the user group of #6 was students who had constraints on time and accessibility on campus. The needs of these user groups were stated as a "means to achieve a healthier lifestyle without spending too much extra time and money". The proposed need of an efficient, healthy lifestyle was well aligned with the busy user group on campus.

Compared to the high parameter value statements, the low parameter statements were less pertinent to the user group because either the user group was too general and not specified enough or the needs were too broad and vague. For the statements like #115 and #119, it was hard to connect the user and needs because the user was "people" or "anyone involved in scientific or technology labs". Like these two statements, either too broad or user groups without any detailed information, hindered the cohesive alignment of user group and their needs. Statement #122 and #124 showed the examples of too vague and broad needs: "To waste less (#122)' and 'to be improved (#124)" lacked adjectives and details to enhance the needs. For the needs of the statement #122, missing details of "what" was wasted and "how much" it should or could be less wasted made the statement less strong. The statement #124 was not only less related to the user group in that it did not provide how the infrastructure(s) could be improved, but also the user, "infrastructure at (The name of University)" was not clarified enough among the broad notion of infrastructure (e.g., system or organization, clinical facilities, offices, centers, communities) (Longtin, 2014).

The high parameter value POV statements identified the user groups' needs and goals in, or with, a verb form so that users could see the choices they could make and choose among the options. In contrast, some of the low parameter value statements' needs provided the needs in a noun form, which described the solution relying on technology, money/funding, a product (specifications), and/or a system (e.g., #117, #118, #119, #120, #121). Although these statements proposed possible solutions, those were limited, predetermined solutions from the perspectives of the writers, not allowing the alternatives from the user's stance. For example, the statement #118 suggested "signage" as a need of their user group to reduce the risk of accidents in the bike lanes. However, this need was a solution and did not include various other possible solutions and the actual needs designers might consider, obviously excluding the possibility that the signage itself might not be the only best solution for the pedestrians.

Another problem found in the low parameter value statements was the interpretation of "need" itself. While most of the high parameter value statements concentrated on the goals and needs user groups experience, some of the low parameter value statements regarded the needs of user groups according to the dictionary definition, as a requirement, necessary duty, or obligation instead of user's goals. This particular type of need misinterpretation can be found in statement #115, #122, and #124. For example, statement #115 highlighted a necessary moral, educational duty of people to be culturally sensitive, statement #122 also emphasized that the user group (food industry) waste less to protect the environment, and statement #124 called for the upgrade of the infrastructure to resolve the overcrowded campus issue. These examples of misinterpretation appeared to affect the insights. Specifically, these misinterpretations appear to lead to a misunderstanding of the problems and current issues specific to the insights for the users.

Insights

A good insight provides the result of meeting the needs, which should be based on the empathy (Gibbons, 2019). It provides the goals user groups can accomplish by solving the current needs, among the multiple possible solutions (Pressman, 2018). In terms of insights, both the high parameter value statements and the low parameter value statements mostly provided the current problem without resolving their current needs, except for statements #2, #3, #5, and #120. These statements provided the positive side the user group could achieve when finding the appropriate solution of the user needs. However, other statements failed to meet this criterion and got high parameter scores regardless of the contents of their insights. For instance, the statement #1 proposed "currently the users lack motivation and different levels of complexity within the class environment" as their insights. However, this was the problem the current situation reveals, not the goal the user group (the school of aviation and transportation technology) are trying to accomplish. The low parameter value statements provided positive goals the user group could achieve but showed the lower parameter value compared to the statement #1. Based on these findings it appeared that, when judging the POV statements, there was a high chance the students did not take the notion of good insights into account. Thus, in terms of insights, the parameter value was not always aligned with the actual quality of the insights.

Summary of the Findings From Construct Validity Analysis

Table 7 provides the summary of the findings from construct validity analysis.

Criterion Validity of Adaptive Comparative Judgment

Concurrent Validity of Adaptive Comparative Judgment

To measure concurrent validity, a correlation was run between the parameter values from conducting the peer reviewed ACJ assessment and the instructors' rubric based grade assignments

TABLE 7 | Summary of findings.

	Highest parameter values	Lowest parameter values
Framework		
Structure and length	- Following basic structures with all necessary components (i.e., user, needs, insights) in a concise manner with necessary details	- Not leading to an actionable statement (e.g., omitted insights) - Include unnecessary information - Short POV statements due to the lack of description
Alignment and logic	- Aligned and actionable	- Lacks alignment, not logical - Not actionable due to the vagueness (e.g., waste less) - Frame the user needs as a specific solution (e.g., implement lights in the parking garage)
Components of POV statements		
User	- Narrowed down with description about the users	- Some of them lacks illustration (e.g., people, pedestrians)
Needs	- Incorporated adjectives and details specific to the user group - Identified the user groups' needs and goals in, or with, a verb form so that users could see the choices	- Less pertinent to the user group because either the user group was too general (e.g., people need to become more educated) - Not specified enough (e.g., Infrastructures need to be improved) - Misinterpretation of 'need' itself (e.g., As a requirement, necessary duty, or obligation instead of user's goals)
Insights	- Both groups showed limitation: parameter value was not always aligned with the actual quality of the insights - Provided the current problem without resolving their current needs (e.g., because it will reduce the carbon footprint that college campuses leave)	

TABLE 8 | Regression results using Assessment 3 (Grades of final deliverable) as the criterion.

Predictor	*b*	*b*	*beta*	*beta*	*sr²*	*sr²*	*r*	Fit
		95% CI [LL, UL]		95%S CI [LL, UL]		95% CI [LL, UL]		
(Intercept)	20.95**	(20.74, 21.16)	—	—	—	—	—	—
Parameter Values	1.50**	(1.13, 1.87)	0.32	(0.24, 0.40)	0.10	(0.06, 0.15)	0.32**	—
—	—	—	—	—	—	—	—	R^2 = 0.101** 95% CI (0.06,0.15)

Note. A significant b*-weight indicates the beta-weight and semi-partial correlation are also significant.* b *represents unstandardized regression weights.* beta *indicates the standardized regression weights.* sr^2 *represents the semi-partial correlation squared.* r *represents the zero-order correlation.* LL *and* UL *indicate the lower and upper limits of a confidence interval, respectively. * indicates* p *< 0.05. ** indicates* p *< 0.01.*

on the POV statements. The peer-evaluated ACJ was not significantly correlated (r = 0.08, p = 0.51) with graders' grading based on rubric. Therefore, the potential concurrent validity of peer-evaluation using ACJ with POV statements is not supported by these results in the context of design thinking.

Predictive Validity of Adaptive Comparative Judgment

As seen in **Table 8**, A simple linear regression was calculated to predict grades of final deliverables (Assessment 3) based on the parameter values of peer-evaluated ACJ (Assessment 1). A significant regression was found (F (1, 575) = 63.057, p < 0.001), with an R^2 of 0.101. Students' predicted grades of final deliverables (Assessment 3) is equal to 20.95 + 1.50 (parameter values). The grades of final deliverables (Assessment 3) increased 1.50 for each point of parameter values of peer-evaluated ACJ (Assessment 1). Therefore, peer-reviewed ACJ showed predictive validity in the context of design thinking.

DISCUSSION

Our research questions guiding the inquiry were: 1) What is the construct validity of ACJ? Does peer-reviewed ACJ reflect general criteria of good POV statements? 2) What is the criterion validity of ACJ? By doing so, this study aimed to validate peer-evaluated ACJ in the design thinking education context. First, this study analyzed ten high parameter value statements and ten low parameter value statements based on the criteria of "good" POV statements (Interaction Design Foundation, 2020; Rikke Friis and Teo Yu, 2020) to examine the construct validity of ACJ. Second, this study examined criterion validity: Concurrent validity and predictive validity. Concurrent validity was studied using correlation between the parameter values and grades on the same POV assignment. Then, the study on the predictive validity was followed to see the parameter values on POV statement can predict future achievement of students, the grades of final deliverables.

The results revealed that peer-evaluated ACJ demonstrated construct validity. The parameter values reflect the quality of POV statements in terms of content structure, needs, user, and insights. The POV statements with higher parameter values showed better quality compared to the POV statements with lower parameter values. This finding is aligned with the findings from previous studies, which reported that ACJ completed by students can be a sound measure for evaluation of self and peer work (Jones and Alcock, 2014; Bartholomew et al., 2020a). Further, the results suggested that peer-evaluated ACJ had predictive validity, but not concurrent validity. When assessing the same POV statements, the results of peer-evaluated ACJ (parameter values) and rubric-based grading by instructors did not show significant correlation. However, the results of peer-evaluated ACJ moderately predicted students' final grades in project 3.

As mentioned in previous studies, peer-evaluated ACJ is not proficient nor professional enough compared to instructors' ACJ (Jones and Alcock, 2014). This may potentially affect the lack of correlation between peer-evaluated ACJ and rubric-based grading of instructors. The lack of correlation between peer evaluated ACJ results and the instructors' rubric based grading may potentially be due to the distributions of the variables as opposed to a lack of concurrent validity. We note that the instructors' rubric based scores are negatively skewed—which we attribute to the criterion-referenced evaluation. Thus, many POV statements may have scored high and similarly to each other on the rubric while in fact there was a noticeable difference between them as discussed in our criterion validity analysis. The ACJ approach yields a norm referenced output which includes a normal distribution regardless of the POV statements meeting the quality standards (or not).

ACJ offers researchers and practitioners in design thinking an effective quality assessment tool that is valid and reliable. As could be seen in the comparison between two groups (i.e., POV statement with high parameter values and POV statements with low parameter values), the results of ACJ displayed the quality of student assignments in a more conspicuous way. The outlier POV statements, such as those generated by teams who failed to progress or high-achiever groups were more notable when using the ACJ, due to its rank system. Early detection of struggling students (or groups) is important for both supporting student's academic achievement in following task and keeping students from dropping out. Instructors could provide timely educational intervention to the student groups who received low parameter values in their task. For instance, if the instructor could support student groups who were struggling in POV statement, he or she could facilitate iteration and revision before student group make a progress using poor-quality POV statement, which might deleteriously affect following design thinking process. Additionally, instructors also could benefit from evaluating the quality of formative assessment during the design projects because goal-oriented, competitive students who were interested in developing one's project in a more excellent manner would be motivated from the results of ACJ.

This study is not without limitations. First, while ACJ provided reliable and valid assessment method, the parameter value highly depends on the relative quality/level of the objects which were being assessed compared. If everyone performs well in the assignment, some students will get low parameter value and rank although the submission successfully meet overall criteria of good POV statements. Therefore, educators should bear the learning objectives and expected outcomes in mind when using ACJ and pay attention to the difference between the higher and lower ranked items. Second, the goal of assessment should be clarified. The rubric based assessment yielded a measure comparing work against a minimum standard where every team could have succeeded. The ACJ measure provided a rank order where one team's POV was strongest, while another weakest. This means that both the strongest and weakest POV's may or may not have met the minimum standards for a good POV statement. Further, peers are students and may not be as proficient as trained graduate students or instructors though they were nearly finished with the course at the time of assessment and the previously-noted work has pointed to the potential for students to complete judgments similarly to experts.

FUTURE IMPLICATIONS

We suspect that an additional benefit of ACJ during the design thinking process was the opportunity for students to learn from both 1) the judgment process and 2) the POV statement examples of their teammates. During the comparative judgment of the POV statements, students had to cognitively internalize criteria to select "better" POV statement and applied those perceptions of quality. Also, the process required students to take a careful look at other students' works as examples of POV statements. Examples resemble the given task and illustrate how the POV-writing task can be completed in the form of near transfer (Eiriksdottir and Catrambone, 2011). Studies revealed that simply being exposed to good examples did not lead to actual transfer (e.g., specify the criteria of good POV statement, explicitly articulate the principles of good POV statement, produce a good POV statement based on what student(s) learn from the POV statements) because learners often do not actively engage in cognitive strategies which help them learning better (Eiriksdottir and Catrambone, 2011). In other words, simply providing good POV examples to the students may not lead to the ability to judge or produce a good POV statement, because students did not use the knowledge from the examples to direct their POV judging/writing process. Educators who were interested in implementing ACJ in the course were required to adopt teaching strategies to enhance transfer of learning from examples such as emphasizing subgoals (Catrambone, 1994; Atkinson et al., 2000) (e.g., articulate main components of POV statements, narrow down the user, set insights as ultimate goal of users), self-explanation (e.g., add detailed explanation about their judging criteria) (Anderson et al., 1997) and group discussion (Olivera and Straus, 2004; Van Blankenstein et al., 2011) (e.g., discuss comparative judgement criteria with peers).

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because data is restricted to use by the investigators as per the IRB agreement. Requests to access the datasets should be directed to nmentzer@purdue.edu.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Purdue University Institutional Review Board.

Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

NM contributed to the research implementation and methodology of this project. WL contributed to the writing, literature review, and statistical analysis. SB contributed to the overall research design and expertise in adaptive comparative judgment.

FUNDING

This material is based on work supported by the National Science Foundation under Grant Number DRL-2101235.

REFERENCES

Anderson, J. R., Fincham, J. M., and Douglass, S. (1997). The Role of Examples and Rules in the Acquisition of a Cognitive Skill. *J. Exp. Psychol. Learn. Mem. Cogn.* 23, 932–945. doi:10.1037//0278-7393.23.4.932

Atkinson, R. K., Derry, S. J., Renkl, A., and Wortham, D. (2000). Learning from Examples: Instructional Principles from the Worked Examples Research. *Rev. Educ. Res.* 70, 181–214. doi:10.3102/00346543070002181

Atman, C. J., Kilgore, D., and McKenna, A. (2008). Characterizing Design Learning: A Mixed-Methods Study of Engineering Designers' Use of Language. *J. Eng. Educ.* 97, 309–326. doi:10.1002/j.2168-9830.2008.tb00981.x

Bartholomew, S. R., Jones, M. D., Hawkins, S. R., and Orton, J. (2021). A Systematized Review of Research with Adaptive Comparative Judgment (ACJ) in Higher Education. *Int. J. Technol. Des. Educ.*, 1–32. doi:10.5296/jet.v9i1.19046

Bartholomew, S. R., Mentzer, N., Jones, M., Sherman, D., and Baniya, S. (2020a). Learning by Evaluating (LbE) through Adaptive Comparative Judgment. *Int. J. Technol. Des. Educ.* 2020, 1–15. doi:10.1007/s10798-020-09639-1

Bartholomew, S. R., Ruesch, E. Y., Hartell, E., and Strimel, G. J. (2020b). Identifying Design Values across Countries through Adaptive Comparative Judgment. *Int. J. Technol. Des. Educ.* 30, 321–347. doi:10.1007/s10798-019-09506-8

Bartholomew, S. R., Strimel, G. J., and Yoshikawa, E. (2019). Using Adaptive Comparative Judgment for Student Formative Feedback and Learning during a Middle School Design Project. *Int. J. Technol. Des. Educ.* 29, 363–385. doi:10.1007/s10798-018-9442-7

Bisson, M.-J., Gilmore, C., Inglis, M., and Jones, I. (2016). Measuring Conceptual Understanding Using Comparative Judgement. *Int. J. Res. Undergrad. Math. Ed.* 2, 141–164. doi:10.1007/s40753-016-0024-3

Borrego, M., Douglas, E. P., and Amelink, C. T. (2009). Quantitative, Qualitative, and Mixed Research Methods in Engineering Education. *J. Eng. Educ.* 98, 53–66. doi:10.1002/j.2168-9830.2009.tb01005.x

Catrambone, R. (1994). Improving Examples to Improve Transfer to Novel Problems. *Mem. Cognit.* 22, 606–615. doi:10.3758/bf03198399

Chapman, V. G., and Inman, M. D. (2009). A Conundrum: Rubrics or Creativity/metacognitive Development? *Educ. Horiz.*, 198–202.

Clemens, N. H., Ragan, K., and Christopher, P. (2018). "Predictive Validity," in *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. Editor B. B. Frey (Thousand Oaks: California: SAGE), 1289–1291.

Coenen, T., Coertjens, L., Vlerick, P., Lesterhuis, M., Mortier, A. V., Donche, V., et al. (2018). An Information System Design Theory for the Comparative Judgement of Competences. *Eur. J. Inf. Syst.* 27, 248–261. doi:10.1080/0960085x.2018.1445461

Dam, R., and Siang, T. (2018). *Design Thinking: Get Started with Prototyping*. Denmark: Interact. Des. Found.

Dochy, F., Gijbels, D., and Segers, M. (2006). "Learning and the Emerging New Assessment Culture," in *Instructional Psychology: Past, Present, and Future Trends*. Editors L Verschaffel, F Dochy, M. Boekaerts, and S. Vosniadou (Amsterdam: Elsevier), 191–206.

Dym, C. L., Agogino, A. M., Eris, O., Frey, D. D., and Leifer, L. J. (2005). Engineering Design Thinking, Teaching, and Learning. *J. Eng. Educ.* 94, 103–120. doi:10.1002/j.2168-9830.2005.tb00832.x

Eiriksdottir, E., and Catrambone, R. (2011). Procedural Instructions, Principles, and Examples: How to Structure Instructions for Procedural Tasks to Enhance Performance, Learning, and Transfer. *Hum. Factors* 53, 749–770. doi:10.1177/0018720811419154

Erickson, J., Lyytinen, K., and Siau, K. (2005). Agile Modeling, Agile Software Development, and Extreme Programming. *J. Database Manag.* 16, 88–100. doi:10.4018/jdm.2005100105

Forman, J., and Damschroder, L. (2007). "Qualitative Content Analysis," in *Empirical Methods For Bioethics: A Primer Advances in Bioethics.* Editors L. Jacoby and L. A. Siminoff (Bingley, UK: Emerald Group Publishing Limited), 39–62. doi:10.1016/S1479-3709(07)11003-7

Gettens, R., Riofrío, J., and Spotts, H. 2015, "Opportunity Thinktank: Laying a Foundation for the Entrepreneurially Minded Engineer." in ASEE Conferences, Seattle, Washington, June 14-17, 2015. doi:10.18260/p.24545

Gettens, R., and Spotts, H. E. (2018). "Workshop: Problem Definition and Concept Ideation, an Active-Learning Approach in a Multi-Disciplinary Setting", in ASEE Conferences, Glassboro, New Jersey, July 24-26, 2018. Available at: https://peer.asee.org/31440.

Gibbons, S. (2019). User Need Statements: The 'Define' Stage in Design Thinking. Available at: https://www.nngroup.com/articles/user-need-statements/.

Haolin, Z., Alicia, B., and Gary, L. (2019). "Full Paper: Assessment of Entrepreneurial Mindset Coverage in an Online First Year Design Course." in 2019 FYEE Conference, Penn State University, Pennsylvania. July 28-30, 2019

Hoge, R. D., and Butcher, R. (1984). Analysis of Teacher Judgments of Pupil Achievement Levels. *J. Educ. Psychol.* 76, 777–781. doi:10.1037/0022-0663.76.5.777

Hsieh, H. F., and Shannon, S. E. (2005). Three Approaches to Qualitative Content Analysis. *Qual. Health Res.* 15, 1277–1288. doi:10.1177/1049732305276687

Hubley, A. M., and Zumbo, B. D. (2011). Validity and the Consequences of Test Interpretation and Use. *Soc. Indic. Res.* 103, 219–230. doi:10.1007/s11205-011-9843-4

Interaction Design Foundation (2020). Point of View - Problem Statement. Available at: https://www.interaction-design.org/literature/topics/problem-statements.

Jones, I., and Alcock, L. (2014). Peer Assessment without Assessment Criteria. *Stud. Higher Edu.* 39, 1774–1787. doi:10.1080/03075079.2013.821974

Jones, I., Swan, M., and Pollitt, A. (2015). Assessing Mathematical Problem Solving Using Comparative Judgement. *Int. J. Sci. Math. Educ.* 13, 151–177. doi:10.1007/s10763-013-9497-6

Jonsson, A., and Svingby, G. (2007). The Use of Scoring Rubrics: Reliability, Validity and Educational Consequences. *Educ. Res. Rev.* 2, 130–144. doi:10.1016/j.edurev.2007.05.002

Karjalainen, J. (2016). "Design Thinking in Teaching: Product Concept Creation in the Devlab Program", European Conference on Innovation and Entrepreneurship, Karjalainen, Janne, September 18, 2016. (Academic Conferences International Limited), 359–364.

Kelley, T. L. (1927). *Interpretation of Educational Measurements*. Oxford, England: World Book Co.

Kernbach, S., and Nabergoj, A. S. (2018). "Visual Design Thinking: Understanding the Role of Knowledge Visualization in the Design Thinking Process", 2018 22nd International Conference Information Visualisation (IV), Fisciano, Italy, July 10-13, 2018 (IEEE), 362–367. doi:10.1109/iv.2018.00068

Kimbell, R. (2008). E-Assessment in Project E-Scape. *Des. Technol. Educ. Int. J.* 12, 66–76.

Lammi, M., and Becker, K. (2013). Engineering Design Thinking. *J. Technol. Educ.* 24, 55–77. doi:10.21061/jte.v24i2.a.5

Lindberg, T., Meinel, C., and Wagner, R. (2010). "Design Thinking: A Fruitful Concept for IT Development?" in *Design Thinking. Understanding Innovation* (Berlin: Springer), 3–18. doi:10.1007/978-3-642-13757-0_1

Longtin, S. E. (2014). Using the College Infrastructure to Support Students on the Autism Spectrum. *J. Postsecond. Educ. Disabil.* 27, 63–72.

Lunz, M. E., and Stahl, J. A. (1990). Judge Consistency and Severity across Grading Periods. *Eval. Health Prof.* 13, 425–444. doi:10.1177/016327879001300405

Lunz, M. E., Wright, B. D., and Linacre, J. M. (1990). Measuring the Impact of Judge Severity on Examination Scores. *Appl. Meas. Edu.* 3, 331–345. doi:10.1207/s15324818ame0304_3

Mahboub, K. C., Portillo, M. B., Liu, Y., and Chandraratna, S. (2004). Measuring and Enhancing Creativity. *Eur. J. Eng. Edu.* 29, 429–436. doi:10.1080/03043790310001658541

Mayring, P. (2015). "Qualitative Content Analysis: Theoretical Background and Procedures," in *Approaches to Qualitative Research in Mathematics Education* (Berlin: Springer), 365–380. doi:10.1007/978-94-017-9181-6_13

Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educ. Res.* 23, 13–23. doi:10.2307/1176219

National Academy of Engineering (2008). Grand Challenges for Engineering. Available at: http://www.engineeringchallenges.org/challenges.aspx.

O'Leary-Kelly, S. W., and Vokurka, R. J. (1998). The Empirical Assessment of Construct Validity. *J. Oper. Manag.* 16, 387–405.

Olivera, F., and Straus, S. G. (2004). Group-to-Individual Transfer of Learning: Cognitive and Social Factors. *Small Group Res.* 35, 440–465. doi:10.1177/1046496404263765

Palisse, J., King, D. M., and MacLean, M. (2021). Comparative Judgement and the Hierarchy of Students' Choice Criteria. *Int. J. Math. Edu. Sci. Tech.*, 1–21. doi:10.1080/0020739x.2021.1962553

Pollitt, A. (2012a). Comparative Judgement for Assessment. *Int. J. Technol. Des. Educ.* 22, 157–170. doi:10.1007/s10798-011-9189-x

Pollitt, A. (2015). On 'Reliability' Bias in ACJ. *Camb. Exam Res.* 10, 1–9. doi:10.13140/RG.2.1.4207.3047

Pollitt, A. (2012b). The Method of Adaptive Comparative Judgement. *Assess. Educ. Principles, Pol. Pract.* 19, 281–300. doi:10.1080/0969594x.2012.665354

Potter, W. J., and Levine-Donnerstein, D. (1999). Rethinking Validity and Reliability in Content Analysis. *J. Appl. Commun. Res.* 27, 258–284. doi:10.1080/00909889909365539

Pressman, A. (2018). *Design Thinking: A Guide to Creative Problem Solving for Everyone*. Oxfordshire: Routledge.

QCDA (1999). Importance of Design and Technology Key Stage 3. Available at: http://archive.teachfind.com/qcda/curriculum.qcda.gov.uk/key-stages-3-and-4/subjects/key-stage-3/design-and-technology/programme-of-study/index.html.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. expanded edition. Chicago: The University of Chicago Press.

Reddy, Y. M., and Andrade, H. (2010). A Review of Rubric Use in Higher Education. *Assess. Eval. Higher Edu.* 35, 435–448. doi:10.1080/02602930902862859

Rikke Friis, D., and Teo Yu, S. (2020). Stage 2 in the Design Thinking Process: Define the Problem and Interpret the Results. Interact. Des. Found. Available at: https://www.interaction-design.org/literature/article/stage-2-in-the-design-thinking-process-define-the-problem-and-interpret-the-results.

Riofrío, J., Gettens, R., Santamaria, A., Keyser, T., Musiak, R., and Spotts, H. (2015, "Innovation to Entrepreneurship in the First Year Engineering Experience." in ASEE Conferences, Seattle, Washington, June 14-17, 2015. doi:10.18260/p.24306

Rowsome, P., Seery, N., and Lane, D. (2013). "The Development of Pre-service Design Educator's Capacity to Make Professional Judgments on Design Capability Using Adaptive Comparative Judgment". in 2013 ASEE Annual Conference & Exposition, Atlanta, Georgia, June 23-26, 2013. 1–10.

Schreier, M. (2012). *Qualitative Content Analysis in Practice*. NY, US. Sage publications.

Seery, N., Canty, D., and Phelan, P. (2012). The Validity and Value of Peer Assessment Using Adaptive Comparative Judgement in Design Driven Practical Education. *Int. J. Technol. Des. Educ.* 22, 205–226. doi:10.1007/s10798-011-9194-0

Sohaib, O., Solanki, H., Dhaliwa, N., Hussain, W., and Asif, M. (2019). Integrating Design Thinking into Extreme Programming. *J. Ambient Intell. Hum. Comput* 10, 2485–2492. doi:10.1007/s12652-018-0932-y

Spooren, P. (2010). On the Credibility of the Judge: A Cross-Classified Multilevel Analysis on Students' Evaluation of Teaching. *Stud. Educ. Eval.* 36, 121–131. doi:10.1016/j.stueduc.2011.02.001

Strimel, G. J., Bartholomew, S. R., Purzer, S., Zhang, L., and Ruesch, E. Y. (2021). Informing Engineering Design Through Adaptive Comparative Judgment. *Eur. J. Eng. Educ.* 46, 227–246.

Thurstone, L. L. (1927). A Law of Comparative Judgment. *Psychol. Rev.* 34, 273–286. doi:10.1037/h0070288

Van Blankenstein, F. M., Dolmans, D. H. J. M., van der Vleuten, C. P. M., and Schmidt, H. G. (2011). Which Cognitive Processes Support Learning during Small-Group Discussion? the Role of Providing Explanations and Listening to Others. *Instr. Sci.* 39, 189–204. doi:10.1007/s11251-009-9124-7

Van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., and De Maeyer, S. (2019). Validity of Comparative Judgement to Assess Academic Writing: Examining Implications of its Holistic Character and Building on a Shared Consensus. *Assess. Educ. Principles, Pol. Pract.* 26, 59–74. doi:10.1080/0969594x.2016.1253542

Wedell-Wedellsborg, T. (2017). Are You Solving the Right Problems. *Harv. Bus. Rev.* 95, 76–83.

Wilson, J., and Wright, C. R. (1993). The Predictive Validity of Student Self-Evaluations, Teachers' Assessments, and Grades for Performance on the Verbal Reasoning and Numerical Ability Scales of the Differential Aptitude Test for a Sample of Secondary School Students Attj7Ending Rural Appalachia Schools. *Educ. Psychol. Meas.* 53, 259–270. doi:10.1177/0013164493053001029

Wolcott, M. D., McLaughlin, J. E., Hubbard, D. K., Rider, T. R., and Umstead, K. (2021). Twelve Tips to Stimulate Creative Problem-Solving with Design Thinking. *Med. Teach.* 43, 501–508. doi:10.1080/0142159X.2020.1807483

Woolery, E. (2019). Design Thinking Handbook. Available at: https://www.designbetter.co/design-thinking April.

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 17 December 2021
doi: 10.3389/feduc.2021.775203

Comparative Judgement for Linking Two Existing Scales

Tom Benton *

Cambridge Assessment, Cambridge, United Kingdom

This article describes an efficient way of using comparative judgement to calibrate scores from different educational assessments against one another (a task often referred to as test linking or equating). The context is distinct from other applications of comparative judgement as there is no need to create a new achievement scale using a Bradley-Terry model (or similar). The proposed method takes advantage of this fact to include evidence from the largest possible number of examples of students' performances on the separate assessments whilst keeping the amount of time required from expert judges as low as possible. The paper describes the method and shows, via simulation, how it achieves greater accuracy than alternative approaches to the use of comparative judgement for test equating or linking.

Keywords: comparative judgement, standard maintaining, equating, linking, assessment

Edited by:
Sven De Maeyer,
University of Antwerp, Belgium

Reviewed by:
San Verhavert,
University of Antwerp, Belgium
Wim Van den Noortgate,
KU Leuven Kulak, Belgium

****Correspondence:***
Tom Benton
benton.t@
cambridgeassessment.org.uk

Specialty section:
This article was submitted to
Assessment, Testing and Applied
Measurement,
a section of the journal
Frontiers in Education

Received: *13 September 2021*
Accepted: *30 November 2021*
Published: *17 December 2021*

Citation:
Benton T (2021) Comparative
Judgement for Linking Two
Existing Scales.
Front. Educ. 6:775203.
doi: 10.3389/feduc.2021.775203

INTRODUCTION

Test equating and linking refers to methods that allow us to identify the scores on one assessment that are equivalent to individual scores on another. This paper concerns the use of comparative judgement (CJ) for linking tests. This context for the use of CJ differs from others in that all the representations included in the CJ study (that is, the exam scripts) already have scores assigned from traditional marking. Therefore, there is no need to use CJ to re-score them. Rather, the aim is simply to calibrate the existing scores from separate assessments onto a common scale. Only enough representations to facilitate calibration need to be included in the associated CJ study. This paper will describe how CJ has been used for test linking in the past, and, more importantly, show how we can improve on existing approaches to increase efficiency.

The idea of using CJ for test linking and equating has existed for a long time. The usual motivation for research in this area is the desire to calibrate assessments from different years against one another. Specifically, to identify grade boundaries on 1 year's test that represent an equivalent level of performance to the grade boundaries that were set on the equivalent test the previous year. A method by which CJ can be used for this task was formalized by Bramley (2005). The method works broadly as described below.

Suppose we have two test versions (version 1 and version 2) and, for each score on version 1, we wish to find an equivalent score on version 2. That is, the score that represents an equivalent level of performance. To begin with, we select a range of representations from each test version. By "representations," for this type of study, we usually mean complete scanned copies of students' responses to an exam paper ("scripts" in the terminology used in British assessment literature). Typically, around 50 representations are selected from each version covering the majority of the score range. Next, the representations are arranged into sets that will be ranked from best to worst by expert judges. In this article, we refer to these sets of representations that will be ranked as "comparison sets" (or just "sets"). In Bramley (2005) and elsewhere these sets of scripts are referred to as "packs." Each comparison set contains representations from both test versions. In a pairwise comparison study, each set would consist of just two representations—one from each test version. For efficiency (particularly in paper-based studies) representations might be arranged into sets of up to 10 each with five representations from version 1 and five from version 2. This process is repeated multiple times (in a paper-based study this involves making multiple physical copies of scripts) so that representations are included in several sets and the precise combination of representations in any set is, as far as possible, never repeated.

When we fit a Bradley-Terry model we are attempting to place all of the representations in the model on a single scale. This process will only work if we have some way of linking every pair of objects in the model to one another by a series of comparisons. For example, representation A may never have been compared to representation B directly. However, if both representation A and representation B have been compared to representations C, D, E and F, then we should be able to infer something about the comparison between representations A and B. The technical term for this requirement is that all objects are *connected*. If our aim is to fit a Bradley-Terry model, then ensuring that all objects are connected to one another is an important part of the design—by which we mean the way in which different representations are assigned to sets (possibly pairs) that will directly compared by judges. Two representations are directly connected if they are ever in the same comparison set. Alternatively, two representations may be indirectly connected if we can find a sequence of direct connections linking one to the other. For example, representations A and D would be indirectly connected if representation A was included in a comparison set with representation B, representation B in a (different) comparison set with representation C, and representation C in (yet another) comparison set with representation D. A design is connected if all possible pairs of representations are connected either directly or indirectly.

Having allocated representations to comparison sets, each set is assigned to one of a panel of expert judges who ranks all of the representations in the set based on their judgements of the relative quality of the performances. In the case of pairwise comparison, where each set consists of only two representations, this simply amounts to the judge choosing which of the two representations they feel demonstrates superior performance.

Once all the representations in each set have been ranked, these rankings are analyzed using a statistical model. For ranking data, the correct approach is to use the Plackett-Luce model (Plackett, 1975), which is equivalent to the rank ordered logit model or exploded logit model described in Allison and Christakis (1994). In the case of pairwise comparisons, analysis is completed using the equivalent, but simpler, Bradley-Terry model (Bradley and Terry, 1952). Whichever model is used, the resulting analysis produces a measure of the holistic quality of each representation depending upon which representations it was deemed superior to, which it was deemed inferior to, and the number of such judgements. These measures of holistic quality (henceforth just "measures") are on a logit scale. This means that, by the definition of the Bradley-Terry model, if representations A and B have estimated CJ measures of θ_A and θ_B, then the probability that a randomly selected judge will deem representation A to display superior performance to representation B (P_{AB}) is given by the equation:

$$P_{AB} = \frac{\exp(\theta_A - \theta_B)}{1 + \exp(\theta_A - \theta_B)}$$

Having fitted a Bradley-Terry model, the performances of all representations are now quantified on a single scale across both test versions. That is, although the test versions are different and the raw scores cannot be assumed to be equivalent, the process of comparative judgement has yielded a single calibrated scale of measures that works across both tests. This can now be used to calibrate the original score scales against one another. The purpose of the final calibration step is that, once it is completed, we can make some inferences about the relative performances of all students that took either of the test versions—not just the sample of students included in the CJ study.

The usual way calibration is completed is illustrated in **Figure 1**. Regression analysis is used to estimate the relationship between scores and measures within each test. Then, the vertical gap between these estimated lines is used to identify the scores on version 2 of the test equivalent to each score on version 1.

Traditionally, the regression lines are not defined to be parallel. However, in most published studies, the differences in the slopes of the two lines are self-evidently small and, on further inspection, usually not statistically significantly different. As a result, in most cases it would make sense to identify a single adjustment figure. That is, how many score points easier or harder is version 2 than version 1? The regression method for this approach would be to identify the most accurate linear predictions of the raw original scores of each representation (denoted x_i for the ith representation) of the form:

$$\hat{x}_i = \beta_0 + \beta_1\theta_i + \beta_2 v_i$$

Where $\hat{x}_i$ is the predicted raw score of the ith representation, θ_i is the CJ measure of the representation, and v_i is a version indicator equal to 1 if the ith representation is from version 2 and equal to 0 otherwise. The coefficients of the regression model are β_0, β_1, and β_2. In this formulation, our particular interest is in the coefficient β_2 which gives a direct estimate of how much easier version 2 is compared to version 1.

The method suggested by Bramley (2005) has been trialed numerous times (e.g., Black and Bramley, 2008; Curcin et al., 2019) and, in general, these trials have produced plausible results regarding the relative difficulty of different test versions.

The regression method above might be labelled score-on-measure as the traditional test scores are the dependent variables and the CJ measures of the quality of each representation are the predictors. However, as described by Bramley and Gill (2010), the regression need not be done this way around. That is, we could perform (measure-on-score) regression with the CJ measures as the dependent variable and the scores as the predictors. Specifically, the regression formula would be:

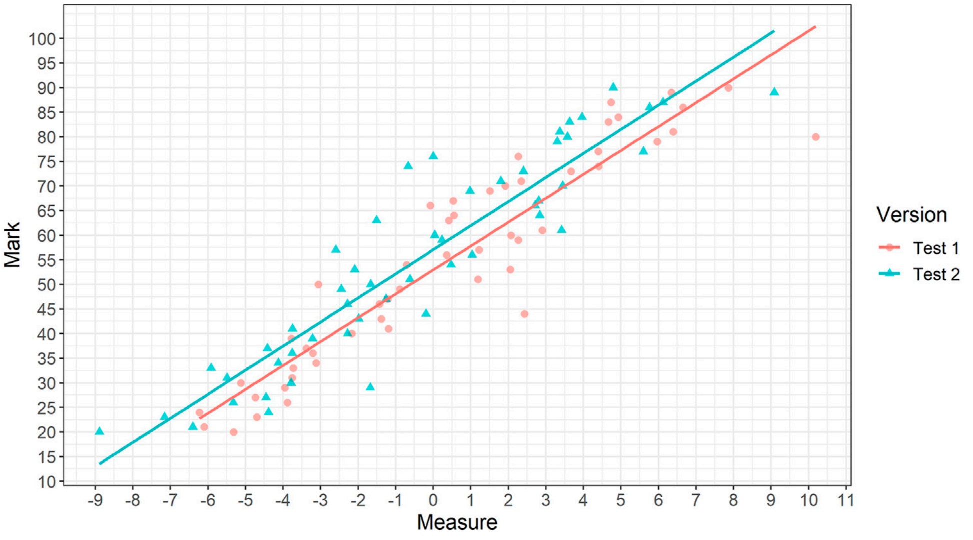


FIGURE 1 | Illustrating the method of linking using CJ suggested by Bramley (2005).

$$\hat{\theta}_i = \gamma_0 + \gamma_1 x_i + \gamma_2 v_i$$

Where $\hat{\theta}_i$ is the predicted CJ measure of the ith representation and γ_0, γ_1, and γ_2 are the regression parameters of this alternative formulation. The relative difficulty of version 2 relative to version 1 is then estimated by $(\frac{-\gamma_2}{\gamma_1})$. In other words, this estimates how much higher a score on version 2 needs to be to predict the same fitted measure as would be predicted by a given score on version 1.

In many practical examples, the differences between the two methods are small (see an investigation by Bramley and Gill, 2010, for one such example). However, during the current research it became clear that large differences between the two methods can occur under certain circumstances. The reasons for this will be explored later in the report. For now, it is sufficient to note that, if the derived CJ measures are reliable and are highly correlated with the original scores then the difference between the two regression approaches should be small. However, knowing that different approaches are possible will be helpful for explaining the results of the simulations later in the report. Other methods of analyzing the same kind of data are also possible (for example, the "standardized major axis," see Bramley and Gill, 2010). However, the two methods mentioned above, along with the new method to be introduced next, are sufficient for the purposes of this paper.

The focus of this paper is to show how a slightly different methodological approach can make the use of CJ for test linking more accurate. In particular, as can be seen from the above description, current approaches to the use of CJ to link existing score scales tend to rely on relatively small samples of representations (around 50) from each test version. Relying on small samples of representations is undesirable as it may lead to high standard errors in the estimates. Since each representation needs to be judged many times by expert judges, under existing approaches, the number of representations included in the study cannot be increased without incurring a significant additional cost. The goal of the newly proposed approach is to allow us to include a greater number of representations in a CJ study to link two existing scales without increasing the amount of time and resource needed from expert judges.

Note that the proposed approach is limited to CJ studies where out goal is to calibrate two existing score scales against one another. As such, the key change in the revised methodology is that it bypasses the need for the Bradley-Terry model in the process. That is, in the newly proposed approach there is no need to conduct a full CJ assessment and produce estimated measures for each representation in the study.

The newly suggested method works as follows. Representations are arranged into pairs of one representation from version 1 of the test and one representation from version 2 of the test. For each pair of representations, an expert judge decides which of the two representations is superior. Next, the difference in scores between the two representations is plotted against whether the representation from version 2 of the test was judged to be superior. An example of such a chart is given in **Figure 2**. The x-axis of this chart denotes the difference in scores. Each judgement is represented by a dot that is close to 1.0 on the y-axis if the version 2 representation is judged superior and is close to 0.0 if it is judged to be inferior (a little jitter has been added to the points to allow them to be seen more easily). As can be seen, in this illustration, where the score awarded to version 2 greatly exceeds the score awarded to version 1, the version 2 representation is nearly always deemed superior. Where the score on version 2 is lower than that on version 1, the version 2 representation is less likely to be judged superior.

The relationship between the score difference and the probability that the version 2 representation is deemed superior is modelled statistically using logistic regression. This is illustrated by the solid blue line in **Figure 2**. To determine how much easier (or harder) version 2 is compared to version 1, the

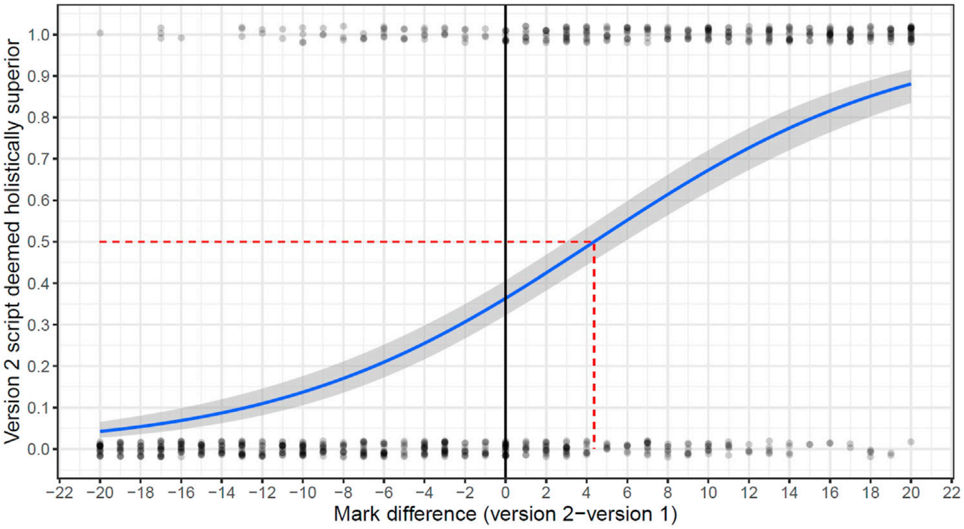


FIGURE 2 | Illustrating the newly proposed method of linking using CJ.

aim is to identify the point at which this line crosses 0.5; that is, where the version 2 representation is as likely to be judged superior as it is to be judged inferior. In the case of **Figure 2**, the data would indicate that version 2 appears to be roughly 4 score points easier than version 1.

We denote the outcome of the ith paired comparison as y_i with a value equal to 1 if the judge deems the representation from test version 2 as superior and equal to zero if the representation from test version 1 is deemed superior. We denote the original score of the representation from test version 2 within the ith paired comparison as x_{2i} and the score of the representation from test version 1 as x_{1i}. The equation for the logistic regression model is as follows.

$$P(y_i = 1) = \frac{\exp(\delta_0 + \delta_1(x_{2i} - x_{1i}))}{1 + \exp(\delta_0 + \delta_1(x_{2i} - x_{1i}))}$$

The δ coefficients in this equation are just the usual logistic regression parameters. The number of score points by which version 2 is easier than version 1 is estimated by $(\frac{-\delta_0}{\delta_1})$.

The newly proposed method, and the avoidance of using a Bradley-Terry model in particular[1], has several advantages:

- There is no need for the same representations to be judged many times. If we were intending to create a reliable set of CJ measures, then it would be necessary for every representation to be judged multiple times. According Verhavert et al. (2019), each representation should be included within between 10 and 14 paired comparisons in order to for CJ measures to have a reliability of at least 0.7. In contrast, the new procedure described above will work even if each representation is only included in a single paired comparison.
- Similarly, because we are not intending to estimate CJ measures for all representations using a Bradley-Terry model, there is no need for the data collection design to be connected.
- As a consequence of the above two advantages, we can include far more representations within data collection without requiring any more time from expert judges. Including a greater number of representations should reduce sampling errors leading to improved accuracy. Whilst in the past, exam scripts were stored physically, they are now usually stored electronically as scanned images. As such, accessing script images is straightforward meaning that the inclusion of greater numbers of representations in a CJ study need not incur any significant additional cost.

Note that, all of the formulae for the new approach can be applied regardless of whether the data collection design collects multiple judgements for each representation, or whether each representation is only included in a single pair. However, we would not expect applying the formulae from the new approach to data that was collected with the intention of fitting a Bradley-Terry model to make estimates any more accurate. The potential for improved accuracy only comes from the fact that the new approach allows us to incorporate greater numbers of representations in a study (at virtually no cost).

We call our new approach to the use of CJ for test linking "simplified pairs." This approach has been described and demonstrated previously in Benton et al. (2020). This current paper will show via simulation, why we expect "simplified pairs"

[1] Of course, we are still using logistic regression and a Bradley-Terry model is itself a form of logistic regression. However, although they can be thought of in this way, Bradley-Terry models usually make use of bespoke algorithms to address issues that can occur in fitting (e.g., see Hunter, 2004). They also have particular requirements in terms of the data collection design (such as connectivity). All of this is avoided.

to provide greater accuracy than previously suggested approaches.

METHODS

A simulation study was used to investigate the potential accuracy of the different approaches to using comparative judgement for linking tests. The parameters for the simulation, such as the specified standard deviation of true CJ measures of the representations and how these are associated with scores, were chosen to give a good match to previous real empirical studies of the use of CJ in awarding. Evidence that this was achieved will be shown as part of the results section.

The process for the simulation study was as follows:

1. Simulate true CJ measures for 20,000 representations from each of test version 1 and test version 2. We denote the true CJ measure of the *i*th representation from test version 1 as θ_i and the true CJ measure of the *j*th representation from test version 2 as θ_j. In both cases these are simulated to follow a normal distribution with a mean of 0 and a standard deviation of 2.
2. Simulate raw scores for the 20,000 representations from each test version. We denote the score of the *i*th representation from test version 1 as x_i and the score of the *j*th representation from test version 2 as x_j. The scores were initially simulated from normal distributions according to

$$x_i = 50 + 8\theta_i + 6\varepsilon_i$$
$$x_j = 54 + 8\theta_j + 6\varepsilon_j$$

Where ε_i and ε_j were simulated from standard normal distributions with a mean of 0 and a standard deviation of 1. These initially simulated scores were rounded to whole numbers and truncated to be between 0 and 100. The resulting simulated scores had means of 50 and 54 for test version 1 and test version 2 respectively. For both test versions, the standard deviation of the resulting scores was approximately 17.

3. Sample 50 representations from version 1 and 50 representations from version 2. Within each test version, sampling was done so that the scores of selected representations were evenly spaced out between 20 and 90.[2]
4. Create the design of a pairwise CJ study that might provide the data for fitting a Bradley-Terry model. This design should ensure that:
   a. Every pair compares a representation from test version 1 to a representation from test version 2.
   b. Each representation is included in N_{CR} pairs (where N_{CR} is a key variable for the study between 2 and 30).
   c. Only representations whose raw scores differ by 20 or less should be paired.
   d. As far as possible, exact pairs of representations are never repeated.

We define T as the total number of pairs in the study. Since we have sample 50 representations from each test version $T = 50{*}N_{CR}$.

5. Simulate the results of the paired comparisons defined in step 4. We imagine that an expert judge has to determine which of the two representations in each pair is superior. The probability that the *j*th representation from version 2 is deemed to display superior performance to the *i*th representation from version 1 is given by the formula:

$$P_{ji} = \frac{\exp\left(\theta_j - \theta_i\right)}{1 + \exp\left(\theta_j - \theta_i\right)}$$

6. Now use the results of this simulated paired comparison study to estimate the difference in difficulty between the two test versions using each of the three methods described earlier. Specifically:
   a. Fit a Bradley-Terry model to the data to generate measures and use a regression of scores on measures.
   b. Based on the CJ estimates from the same Bradley-Terry model, use a regression of measures on scores.
   c. Directly estimate the difference in difficulty between test versions using the logistic regression method described earlier. This represents using the analysis methodology from our newly suggested approach but without taking advantage of the potential improvements to the data collection design.
7. Now, using the same set of 20,000 representations from each version (from steps 1 and 2), simulate a full simplified pairs study. The aim is that the study will include the same number of pairs as the other methods (i.e., T), but that we will sample more representations and only include each of them in a single pair. To begin with, we sample T representations from version 1 and T representations from version 2. Within each test version, representations were again selected so that their scores were evenly spaced out between 20 and 90.
8. Using these freshly selected representations, create the design of a simplified pairs study (i.e., assign representations to pairs). This design should ensure that:
   a. Every pair compares a representation from test version 1 to a representation from test version 2.
   b. Each representation is included in exactly 1 pair.
   c. Only representations whose raw scores differ by 20 or less should be paired.

Since each representation is included in a single pair this will result in T pairs.

9. Simulate the results of these fresh paired comparisons using the same formula as in step 5.
10. Using the data from these fresh paired comparisons, apply logistic regression to generate an estimate of the relative difficulty of version 1 and version 2. This is the simplified

[2] An even spread of 50 values between 20 and 90 is first defined by the sequence of number 20.00, 21.43, 22.86, 24.29,. . ., 88.57, 90.00. For each of these values in turn, we randomly select one script from those with raw scores as close as possible to these values. That is, from those with raw scores of 20, 21, 23, 24,. . .,89, 90.

pairs estimate of the difference in the difficulty of the two tests.

11. Repeat the entire process (steps 1–10) 2,000 times.

All analysis was done using R version 4.0.0 and the Bradley-Terry models were fitted using the R package *sirt* (Robitzsch, 2019).

The above procedure was repeated with the total number of pairs in each study (T) taking each of the values 100, 200, 300, 400, 500, 750, 1000, and 1500. For every method other than the full simplified pairs approach, where each representation is only included in a single paired comparison, these values correspond to the number of paired comparisons for each of the 50 representations for each test version (N_{CR}) being 2, 4, 6, 8, 10, 15, 20, and 30.

Note that the first 2 steps of the simulation process produce realistic means and standard deviations of the simulated scores. That is, the means (50 and 54 for the two respective test versions) and the standard deviations (approximately 17 for each test version) are typical of the values we tend to find in real tests of this length.

As can be seen from the above description, test version 2 is simulated to be exactly 4 score points easier than version 1. This size of difference in difficulty was chosen as it reflects the typical absolute amount (as the percentage of maximum available score) by which GCSE component grade boundaries changed between 2015 and 2016[3]. As such, it is typical of the kind of difference we'd need our methods to handle in practice.

As mentioned above, the way in which representations were sampled to be evenly spread across the score range from 20 to 90 per cent (steps 3 and 7) reflects the way previous CJ studies for linking tests have been done in practice. Representations with very high scores are usually excluded as, if two candidates have answered nearly perfectly, it can be extremely difficult to choose between them. Representations with scores below 20 per cent of the maximum available are also typically excluded as, in practice, they often have many omitted responses meaning that judges would have very little evidence to base their decisions on.

Further evidence of how the simulation design produces results that are representative of real studies of this type will be provided later.

The aim of analysis was to explore the accuracy with which each of the different methods correctly identified the true difference in difficulty between the two test versions (4 score points). This was explored both in terms of the bias of each method (i.e., the mean estimated difference across simulations compared to the true difference of 4), and the stability of estimated differences across simulations.

In addition to recording the estimated differences in difficulty using each method within each simulation, we also recorded the standard errors of the estimates that would be calculated for each method. This helps to understand how accurately each method would allow users to evaluate the precision of their estimates. Specifically:

- For the score-on-measure regression approach the standard error of the estimated difference in difficulty is simply given by the standard error of β_2 in the regression. Note that the use of this standard error requires that the assumptions underpinning the regression itself are correct. However, the usual assumption that the observations in the regression are independent (e.g., in **Figure 1**) is, in fact, incorrect. Since all CJ measures were estimated simultaneously, the CJ measures of different representations are, in fact, correlated with larger (positive) correlations between representations that were directly compared. Despite this concern, these kind of estimates of uncertainty have been used in previous research (e.g., Curcin et al., 2019) and it was of interest to examine their accuracy.
- For the measures on scores approach the standard error of any estimate is derived using the delta method. Specifically, if we label the parameter covariance matrix from the regression model as $C(\boldsymbol{\gamma})$, then the standard error of the estimated difference in difficulty is given by:

$$Standard\ Error = \sqrt{G^T C(\boldsymbol{\gamma}) G}, \text{where } G = \begin{pmatrix} 0 \\ \left(\frac{\gamma_2}{\gamma_1^2}\right) \\ \left(\frac{-1}{\gamma_1}\right) \end{pmatrix}$$

Once again, these standard errors rely on the assumptions of the regression being correct and, as such, may suffer from the same issues as those based on scores on measures regression.

- For the simplified pairs method, we can also use the delta method to create standard errors. Specifically, if we denote the parameter covariance matrix from the logistic regression as $V(\boldsymbol{\delta})$ then the standard error of the estimated difference in difficulty is given by:

$$Standard\ Error = \sqrt{H^T V(\boldsymbol{\delta}) H}, \text{Where } H = \begin{pmatrix} \left(\frac{-1}{\delta_1}\right) \\ \left(\frac{\delta_0}{\delta_1^2}\right) \end{pmatrix}$$

These standard errors rely on the assumptions underpinning the logistic regression being correct. Within our simulation these assumptions are plausible for the full simplified pairs approach. In particular, if each representation is only used once,

[3] GCSE stands for General Certificate of Education. GCSEs are high-stakes examinations taken each summer by (nearly) all 16-year-olds in England and OCR is one provider of these examinations. The years 2015 and 2016 were chosen as they were comfortably after the previous set of GCSE reforms and the last year before the next set of GCSE reforms began. As such, they represented the most stable possible pair of years for analysis. Only grades A and C were explored and only examinations that were taken by at least 500 candidates in each year. At grade A the median absolute change in boundaries was 3.8 per cent of marks. At grade C the median absolute change in boundaries was 3.3 per cent of marks.

observations in the logistic regression are independent. For the use of the logistic regression approach based on the simulated data where the same representations are used multiple times the assumption of the independence of observations is quite clearly incorrect and so these standard errors were not retained[4].

To help verify the realistic nature of the simulation study, for all methods using a Bradley-Terry model, the reliability of the CJ measures was recorded within each simulation. This was calculated both in terms of an estimated scale separation reliability (SSR, see Bramley, 2015) and also in terms of the true reliability calculated as the squared correlation between estimated CJ measures and the true simulated values. Correlations between estimated CJ measures and raw scores on each test version were also calculated and recorded from each simulation.

RESULTS

A Digression on the Realistic Nature of the Simulation

To begin with it is worth noting that, by design, the simulation produced results regarding the reliability of CJ measures that were very consistent with those typically seen in empirical studies. For example, for the simulations involving 750 comparisons in total and 15 per representation (a typical number of comparisons per representation in previous studies of this type), across simulations, the median SSR was 0.93 (the median true reliability[5] was also 0.93), and the median correlation between CJ measures and raw scores was 0.92 (for both test versions). These values match the median reliabilities and correlations between raw scores and estimated CJ measures across 10 real studies based on using pairwise comparative judgement to link score scales published by Curcin et al. (2019, page 41, Table 7).

The average level of reliability from 15 comparisons per representation (0.93), which matches the average values from real empirical studies of this type (Curcin et al., 2019), is somewhat higher than research on the use of CJ in other contexts suggests is typical (for example, see Verhavert et al., 2019). Although this is not the main focus of the article, we will briefly digress to explain why the discrepancy occurs. In short, we believe it is largely because, in studies concerned with linking two existing scales, all representations have already been scored in a non-CJ way to begin with. The analysis can capitalize on this additional data from the original scores in ways that are not possible if CJ if the sole method by which representations are being assessed. Of course, there is a cost to scoring all representations before beginning a CJ study, so this should not be taken as a recommendation that this should be done in general.

Part of the reason for the higher reliability coefficients in empirical CJ studies concerned with linking existing scales (e.g., Curcin et al., 2019) is the way in which representations are selected. Unlike the studies by Verhavert et al. (2019), only a sample of the possible representations are included in the CJ study and this sample is not selected at random. Rather, representations are deliberately selected with scores that are evenly spread across the available range between 20 and 90 per cent of the paper total. This ensures that a wider range of performances is included in each study than would be the case by selecting representations purely at random. We would expect this to mean that the standard deviation of the true CJ measures included in such a study is higher than in the population in general and, as a result, reliability coefficients are expected to be higher.

In addition, because, by design, representations are only compared to those with relatively similar scores, some of the advantages usually associated with adaptive comparative judgement (ACJ, see Pollitt, 2012) are built into the method. This allows higher reliabilities to be achieved with smaller numbers of comparisons. Note that, although the method has some of the advantages of ACJ, it is not actually adaptive. Which representations are compared to one another is not amended adaptively dependent upon the results of previous comparisons. As such, concerns about the inflation of reliability coefficients in an adaptive setting (Bramley and Vitello, 2019) do not apply.

Understanding the reasons for these high reliability coefficients, and that these reflect the values that we see on average in real empirical studies of this type is important as it allows us to have confidence in the remainder of the results presented in this paper.

Before returning to the main subject of this paper we note that, as expected, within our own simulation study, the reliability of the CJ measures increased with the number of comparisons per representation. The median reliability was just 0.2 if only 2 comparisons per representation were used, rose to above 0.7 for 4 comparisons per representation, and was 0.96 for 30 comparisons per representation[6].

Biases and Standard Errors of Different Methods

Our main interest is in the bias and variance (i.e., stability) of the various methods for estimating the relative difficulty of two tests. **Figure 3** shows the results of the analysis in terms of the mean estimated difference in the difficulty of the two tests from each method. The mean estimated difference from each method is compared to the known true difference (4 score points) represented by the thick grey line. The mean difference between the estimated and actual differences in test difficulty provides an estimate of bias and this is shown by the secondary y-axis on the right-hand side. Note that the method labelled

[4] It is possible to address this issue via the application of multilevel modelling (see Benton et al., 2020, for details). However, this changes the estimates themselves and is beyond the scope of this article so was not considered here.

[5] True reliabilities are calculated as the squared correlation between estimated CJ measures and the true values of CJ measures (i.e., simulated values).

[6] Based on true reliabilities. Note that true reliabilities and scale separation reliabilities were always very close to one another except where the number of comparisons per script was below 5.

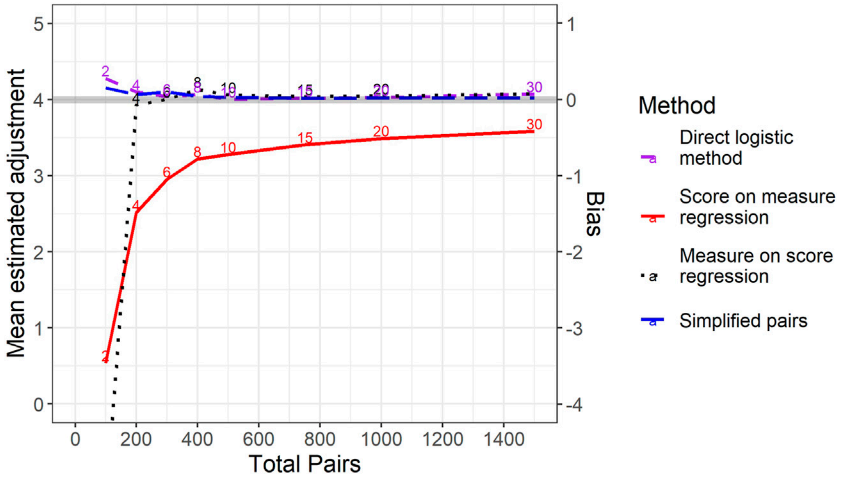


FIGURE 3 | Mean estimated difference in difficulty between test versions across simulations for different methods by total number of pairs per study. Note that the true level of difference in difficulty is 4 (the solid grey line). For the three methods in which representations were included in multiple pairs, the number of pairs per representation is noted just above the relevant line.

"direct logistic method" (the purple dashed line) relates to results from applying the newly proposed analysis methodology but with the same data as other methods. In contrast the method labelled "simplified pairs" relates to using our method from a data collection design with each representation being included in just 1 comparison. As long as the total number of pairs in the study is at least 200, 3 of the methods have levels of bias very close to zero. The two approaches based on logistic regression are essentially unbiased across all sample sizes. However, the most interesting result from **Figure 3** is the evident bias of using the Bradley-Terry model in combination with a regression of scores on measures—that is, the dominant method of using comparative judgement in standard maintaining in studies of this type to date.

The score-on-measure regression method has a negative bias. That is, on average it underestimates the scale of the difference in difficulty between the two test versions. The reason for this is to do with the way in which representations are selected for most studies of this type. To understand why this is, imagine a situation where, perhaps due to having a very small number of comparisons per representation, the CJ measure was utterly unreliable and had zero correlation with the scores awarded to representations. In this instance, the score-on-measure regression (e.g., **Figure 1**) would yield two horizontal lines. The vertical distance between these lines would actually be pre-determined by the difference in the mean scores of the representations we had selected from each version. In our study, since we have deliberately selected the same range of scores for each test version, this difference is equal to zero.

As the number of comparisons per representation increases, the size of the bias reduces but does not immediately disappear. With low, but non-zero, correlations between scores and measures the estimated difference between test versions will hardly be adjusted from the (predetermined) mean difference between the selected representations. As such, the bias in the method would persist. As the number of comparisons per representation increases, this bias becomes much smaller. However, due to the fact that, in this simulation, even the true CJ measures are not perfectly correlated with scores (correlation of 0.95) this bias never completely disappears.

Aside from bias, we are also interested in the stability of estimates from different methods—that is, their standard errors. According to the Cambridge Dictionary of Statistics (Everitt and Skrondal, 2010) a standard error (SE) is "the standard deviation of the sampling distribution of a statistic" (page 409). In our case, the "statistic" we are interested in is the estimated difference in the difficulty of two tests by some method and the "sampling distribution" is observable from the simulations we have run. As such, we can calculate the true standard error of each method by calculating the standard deviation of estimated differences in difficulty across simulations. **Figure 4** shows how the standard errors of the estimates of differences in difficulty change depending upon the total number of pairs in the study. Somewhat counterintuitively, **Figure 4** shows that the score-on-measure approach becomes less stable (i.e., has higher standard errors) as the number of comparisons per representation increases. In other words, increasing the amount of data we collect makes the results from this method more variable. This result is due to the fact that, as described above, where the correlation between original raw scores and measures is low, the method will hardly adjust the estimated difference in the difficulty of test versions from the predetermined mean score difference of zero. As such, across multiple replications of the simulations with low numbers of comparisons per representation, the score-on-measure method will reliably give an (incorrect) estimate close to zero. As the number of comparisons per representation increases, and the correlation between scores and measures becomes stronger, so the method will actually begin making substantive adjustments to account for differences in the holistic quality of responses and so the results become more variable across simulations.

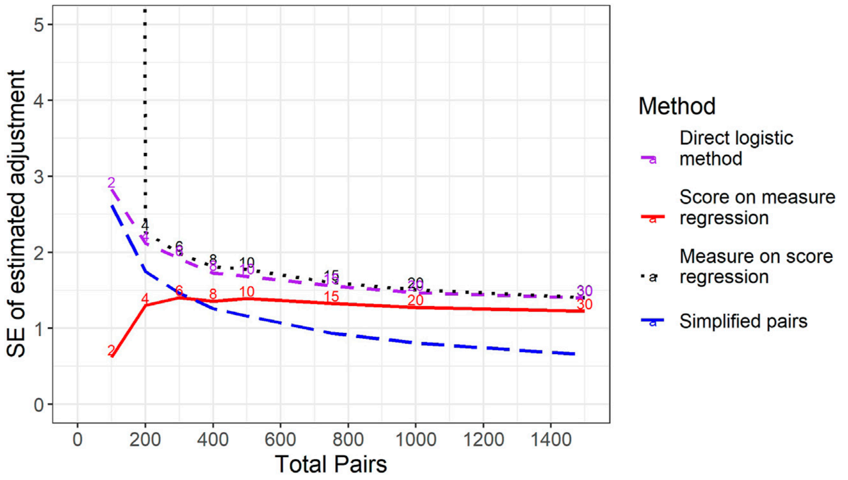


FIGURE 4 | Standard deviation of estimated difference in difficulty between test versions across simulations for different methods by total number of pairs per study. For the three methods in which representations were included in multiple pairs, the number of pairs per representation is noted just above the relevant line.

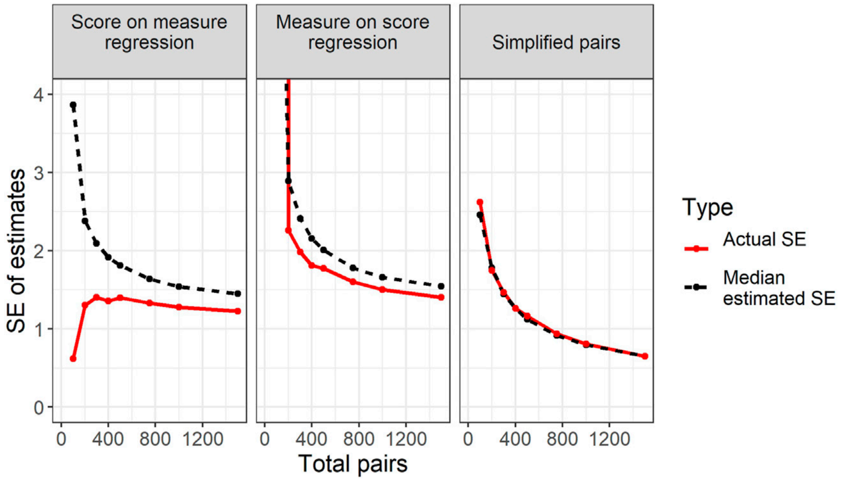


FIGURE 5 | Plot of median estimated standard errors of each method and actual standard deviations of estimated difference in difficulty for different total study sizes.

Figure 4 shows that if we only allow 2 comparisons per script then the measure-on-score regression approach is extremely unstable. However, for larger numbers of comparisons per script, the standard errors of the measure-on-score and the direct logistic methods using the same set of data are very similar. This is, perhaps, unsurprising as, in essence, both methods are doing the same thing, although the score-on-measure regression uses one more step in the calculation. That is, both methods attempt to find the score difference where representations from either test version are equally likely to be deemed superior by a judge.

Of most interest are the simplified pairs results based on using the same total number of paired comparisons but only using each representation once. For any given number of total pairs, this approach is more stable than either of the two alternative unbiased methods (measure-on-score regression or direct logistic). Furthermore, the simplified pairs approach yields roughly the same standard errors with 300 comparisons in total as can be achieved with five times as many comparisons (30 per representation or 1500 in total) for either of the other two approaches. This suggests that avoiding the use of the Bradley-Terry model, including as many different representations as possible in the exercise, and using logistic regression to estimate the difference in the difficulty of two test versions can lead to huge improvements in efficiency in terms of the amount of time required from expert judges. This also suggests that including 300 comparisons in a simplified pairs study should provide an acceptable level of reliability.

Figure 4 concerns the true standard errors of each method – that is, the actual standard deviations of estimates across

TABLE 1 | Coverage probabilities for three methods dependent upon the total number of pairs in the study.

Total pairs in study	Coverage probability for score-on-measure regression (%)	Coverage probability for measure-on-score regression (%)	Coverage probability for simplified pairs (%)
100	100.0	100.0	95.8
200	99.0	99.2	96.0
300	98.0	98.9	95.6
400	98.2	98.6	95.5
500	97.3	97.9	95.1
750	97.1	97.6	95.2
1000	97.1	96.7	94.3
1500	97.3	97.0	95.8

simulations. However, true standard errors are not generally observable outside of simulation studies and we need an alternative way of estimating standard errors in practice. How this might be done within each approach was described earlier and some formulae were provided. **Figure 5** compares the median estimated standard errors of each method, based on the formulae provided earlier, to actual standard errors within each size of study. As can be seen, estimated standard errors for both score-on-measure and measure-on-score regression tend to be too high. The reasons for this as regards score-on-measure regression have largely already been discussed. For measure-on-score regression the issue relates to the assumptions of the regression model.

The estimated standard errors come from the regression of CJ measures on scores using data of the type shown in **Figure 1**. Estimating standard errors essentially involves asking how much we'd expect the gap between two regression lines to change if we were to rerun the study with a fresh sample of representations. In some studies, though not here, this is estimated using bootstrapping (e.g., Curcin et al., 2019) which involves literally resampling from the points in charts like **Figure 1** (with replacement) many times and measuring the amount by which the gap between lines varies.

The fact that the assumption of independent errors does not hold, explains the discrepancy between the actual and estimated standard errors of measure-on-score regression. Specifically, because every comparison is between a version 1 representation and a version 2 representation, the gap between regression lines will be less variable across samples than would be expected by imagining every point in the regression as being independent. In short, ensuring that every comparison in a pairwise design is between versions is a good thing because it reduces the instability of the gap between regression lines. However, it is a bad thing for accurately estimating standard errors as it leads to a violation of the regression assumptions.

In the simulations described here, estimated confidence intervals based purely on the regression chart tend to be wider than necessary. In other situations, we would expect the error in estimation to work the other way. For example, imagine that the design of a CJ study included large numbers of comparisons within test versions but only a handful of comparisons between version 1 and version 2. Instinctively, we can tell that such a design would provide a very poor idea of the relative difficulty of the two test versions. However, with sufficient comparisons within versions, we could generate high reliability statistics, and high correlations between scores and measures within versions. As such, we could produce a regression chart like **Figure 1** that appeared reassuring. In this case, confidence intervals based on the data in the regression alone would be far too narrow and would not reflect the true uncertainty in estimates.

Regardless of the reasons, the importance of the findings here is to show that not only is the simplified pairs method unbiased and more stable than alternative approaches, it is also the only method where we can produce trustworthy estimates of accuracy through standard errors. This is further shown by **Table 1**. This table shows the coverage probabilities of the 95% confidence intervals for each method. These confidence intervals are simply calculated to be each method's estimate of the difference in difficulty between test versions plus or minus 1.96 times the estimated standard error. **Table 1** shows the proportion of simulations of each size (out of 2000) where the confidence interval contains the true difference in difficulty (4 score points). For both regression-based approaches, the coverage probabilities are substantially higher than the nominal levels confirming that the estimated standard errors tend to be too high. However, for simplified pairs the coverage probabilities are close to the intended nominal level.

Unlike the other CJ approaches, in the simplified pairs method, we are not attempting to assign CJ measures to representations. As such, we do not calculate any reliability coefficients analogous to the SSR. Rather, the chief way in which we assess the reliability of a simplified pairs study in practice is by looking at the estimated standard errors. With this in mind, it is reassuring that the analysis here suggests we can estimate these accurately.

CONCLUSION

This paper has reviewed some possible approaches to using expert judgement to equate test versions. In particular, the research has evaluated a new approach (simplified pairs) to this problem and shown via simulation that we expect it to be more efficient than existing alternatives, such as that suggested by Bramley (2005), that rely upon the Bradley-Terry model. Improved efficiency is possible because, by changing the way results are analyzed, we can include a far higher number of representations within data collection without increasing the workload for judges. The simplified pairs approach is also the only approach where we can produce trustworthy confidence intervals for the estimated relative difficulty of two tests.

The analysis has also revealed some weaknesses in the traditional approach based on regression of the scores awarded to representations on measures of holistic quality from a CJ study. In particular, the results indicate that this method is biased towards the difference in the mean scores of the representations selected for the study. Given that the whole point of analysis is to provide fully independent evidence of the relative difficulty of two tests, such biases are undesirable.

The results in this paper suggest that, using a simplified pairs approach, a CJ study based on no more than 300 paired comparisons in total may be sufficient to link scores scales across test versions reasonably accurately. It is worth considering how this workload compares to a more traditional awarding meeting (not based on CJ) where expert judges would attempt to set grade boundaries on 1 year's exam that maintain standards from previous years. According to Robinson (2007), in the past, traditional awarding meetings in England would generally involve at least eight examiners. In these meetings, each examiner would be expected to review at least seven exam scripts within a range of plus or minus three from a preliminary grade boundary. This process might be repeated for up to three separate grade boundaries (for example, grades A, C and F in England's GCSE examinations). Thus, a total of 168 (=8 judges*7 scripts per grade*3 grades) script reviews might have taken place within an awarding meeting. With this in mind, it is clear that the current suggestion of a CJ-based process requiring 300 paired comparisons would require more resources than traditional awarding—although not of a vastly increased order of magnitude.

It is worth noting that the suggested method, based on logistic regression, does require a few assumptions. In particular, the suggested logistic regression method assumes a linear relationship between the difference in the raw scores of the representations being compared and the log odds of the representation from a particular test version being judged superior. In addition, the method assumes that the relationship between score differences and judged representation superiority is constant across all of the judges in a study. In practice, both of these assumptions could be tested using the grouping method described in chapter 5 of Hosmer and Lemeshow (2000). If there was any sign of lack of fit then it is fairly straightforward to adjust the model accordingly, for example, by adding additional (non-linear) terms to the logistic regression equation. If there were evidence that results varied between different judges, then it would be possible to use multilevel logistic regression as an alternative with judgements nested within judges to account for this.

This paper has only provided detailed results from one simulation study. However, it is fairly easy to generalize the results to simulations with different parameters. For example:

- We know that the score-on-measure regression method is biased towards the difference in the mean scores of sampled representations from different test versions (zero in our study). As a result, the greater the true difference in difficulty between test versions, the greater the level of bias we'd expect to see.
- By the same logic, if representations were randomly sampled rather than selected to be evenly spaced over the range of available scores, then the mark-on-measure regression method would be biased towards the difference in population means rather than towards zero. In our simulated example this would be an advantage. However, in practice, due to the changing nature of students entering exams in different years the difference in population means may or may not reflect the difference in the difficulty of the two tests. One change from the earlier results would be that, due to random sampling, the standard deviation of estimated differences *via* score-on-measure regression (e.g., **Figure 4**) would decrease rather increase with the number of pairs in the study.
- It is also fairly easy to predict the impact on results of reducing the spread of true CJ measures in the simulation. This naturally leads to the estimated CJ measures being less reliable. With estimated CJ measures being less reliable, the bias of the score-on-measure regression method would increase. Aside from this, the reduced reliability of all CJ measures would reduce the stability of all other methods. This includes simplified pairs where the reduced spread of true CJ measures would lead to a weakening of the relationship between score differences and the decisions made by judges – in turn leading to reduced stability in estimates.

Although, for brevity, results are not included in this paper, the suggestions in the above bullets have all been confirmed by further simulations. Whilst it is possible to rerun our simulation with different parameters it is worth noting that the parameters of the simulation presented in this paper have been very carefully chosen to reflect a typical situation that is likely to be encountered in practice. As such, the results that have been presented provide a reasonable picture of the level of accuracy that can be achieved *via* the use of CJ for linking or equating.

Aside from simulation, demonstrations of the simplified pairs technique in practice can be found in Benton et al. (2020). This includes details on how the method can be extended to allow the difference in the difficulty of two tests to vary across the score range. The combination of theoretical work based on simulation (this current paper) and previous empirical experimental work indicate that simplified pairs provides a promising mechanism by which CJ can inform linking and equating.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Allison, P. D., and Christakis, N. A. (1994). Logit Models for Sets of Ranked Items. *Sociological Methodol.* 24, 199–228. doi:10.2307/270983

Benton, T., Cunningham, E., Hughes, S., and Leech, T. (2020). *Comparing the Simplified Pairs Method of Standard Maintaining to Statistical equatingCambridge Assessment Research Report*. Cambridge, UK: Cambridge Assessment.

Black, B., and Bramley, T. (2008). Investigating a Judgemental Rank-ordering Method for Maintaining Standards in UK Examinations. *Res. Pap. Edu.* 23 (3), 357–373. doi:10.1080/02671520701755440

Bradley, R. A., and Terry, M. E. (1952). Rank Analysis of Incomplete Block Designs: The Method of Paired Comparisons. *Biometrika* 39, 324–345. doi:10.1093/biomet/39.3-4.324

Bramley, T. (2005). A Rank-Ordering Method for Equating Tests by Expert Judgment. *J. Appl. Meas.* 6, 202–223.

Bramley, T., and Gill, T. (2010). Evaluating the Rank-ordering Method for Standard Maintaining. *Res. Pap. Edu.* 25 (3), 293–317. doi:10.1080/02671522.2010.498147

Bramley, T. (2015). *Investigating the Reliability of Adaptive Comparative JudgmentCambridge Assessment Research Report*. Cambridge, UK: Cambridge Assessment.

Bramley, T., and Vitello, S. (2019). The Effect of Adaptivity on the Reliability Coefficient in Adaptive Comparative Judgement. *Assess. Educ.: Princ. Policy Pract.* 26 (1), 43–58.

Curcin, M., Howard, E., Sully, K., and Black, B. (2019). Improving Awarding: 2018/2019 Pilots. Ofqual Report Ofqual/19/6575. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/851778/Improving_awarding_-_FINAL196575.pdf (Accessed August 11, 2021).

Everitt, B. S., and Skrondal, A. (2020). *The Cambridge Dictionary of Statistics*. 4th Edn., Cambridge University Press.

Hosmer, D. W., and Lemeshow, S. (2000). *Applied Logistic Regression*. New York: John Wiley & Sons.

Hunter, D. R. (2004). MM Algorithms for Generalized Bradley-Terry Models. *Ann. Stat.* 32, 384–406. doi:10.1214/aos/1079120141

Plackett, R. L. (1975). The Analysis of Permutations. *Appl. Stat.* 24, 193–202. doi:10.2307/2346567

Pollitt, A. (2012). The Method of Adaptive Comparative Judgement. *Assess. Educ. Principles, Pol. Pract.* 19 (3), 281–300. doi:10.1080/0969594x.2012.665354

Robinson, C. (2007). "Awarding Examination Grades: "Current Processes and Their Evolution," in *Techniques for Monitoring the Comparability of Examination Standards*. Editors P. E. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (London: Qualifications and Curriculum Authority), 97–123.

Robitzsch, A. (2019). sirt: Supplementary Item Response Theory Models. R Package Version 3.7-40. Available at: https://CRAN.R-project.org/package=sirt.

Verhavert, S., Bouwer, R., Donche, V., and De Maeyer, S. (2019). A Meta-Analysis on the Reliability of Comparative Judgement. *Assess. Educ. Principles, Pol. Pract.* 26, 541–562. doi:10.1080/0969594x.2019.1602027

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Education

ORIGINAL RESEARCH
published: 09 February 2022
doi: 10.3389/feduc.2021.785919

The Accuracy and Efficiency of a Reference-Based Adaptive Selection Algorithm for Comparative Judgment

San Verhavert[1]*, *Antony Furlong*[2] and *Renske Bouwer*[1,3]

[1]Department of Training and Education Sciences, University of Antwerp, Antwerp, Belgium, [2]International Baccalaureate (IB), The Hague, Netherlands, [3]Utrecht Institute of Linguistics OTS, Utrecht University, Utrecht, Netherlands

Several studies have proven that comparative judgment (CJ) is a reliable and valid assessment method for a variety of competences, expert assessment, and peer assessment, and CJ is emerging as a possible approach to help maintain standards over time. For consecutive pairs of student works (representations) assessors are asked to judge which representation is better. It has been shown that random construction of pairs leads to very inefficient assessments, requiring a lot of pairwise comparisons to reach reliable results. Some adaptive selection algorithms using information from previous comparisons were proposed to increase the efficiency of CJ. These adaptive algorithms appear however to artificially inflate the reliability of CJ results through increasing the spread of the results. The current article proposes a new adaptive selection algorithm using a previously calibrated reference set. Using a reference set should eliminate the reliability inflation. In a real assessment, using reference sets of different reliability, and in a simulation study, it is proven that this adaptive selection algorithm is more efficient without reducing the accuracy of the results and without increasing the standard deviation of the assessment results. As a consequence, a reference-based adaptive selection algorithm produces high and correct reliability values in an efficient manner.

Keywords: comparative judgment, assessment, adaptive selection algorithm, adaptive, efficiency, reliability, accuracy, reference set

OPEN ACCESS

Edited by:
Anders Jönsson,
Kristianstad University, Sweden

Reviewed by:
Jeffrey Buckley,
Athlone Institute of Technology,
Ireland
Tom Bramley,
Cambridge Assessment,
United Kingdom

***Correspondence:**
San Verhavert
san.verhavert@uantwerpen.be

Specialty section:
This article was submitted to
Assessment, Testing and Applied
Measurement,
a section of the journal
Frontiers in Education

Received: *29 September 2021*
Accepted: *14 December 2021*
Published: *09 February 2022*

Citation:
Verhavert S, Furlong A and Bouwer R
(2022) The Accuracy and Efficiency of
a Reference-Based Adaptive Selection
Algorithm for Comparative Judgment.
Front. Educ. 6:785919.
doi: 10.3389/feduc.2021.785919

INTRODUCTION

Comparative judgment (CJ) is a recent, alternative form of assessment. A group of assessors are individually presented with several, consecutive pairs of works of students (hereafter called representations). For every pair, the assessors are asked which of the two representations is better considering the task or competency under assessment. Based on the pairwise judgments of the assessors, logit scores can be estimated using the Bradley–Terry–Luce (BTL) model (Bradley and Terry 1952; Luce 1959). These logit scores, also referred to as ability scores, represent the consensus view of the group of assessors about the quality of the representations in regard of the task or competency under assessment. Specifically, a logit score represents the difference in quality (in a log-transformed probability) between a particular representation and a representation of average quality for this group of representations. The strength of this method is based on the observation that in everyday life all judgments a person makes are in fact comparisons (Laming 2003). Furthermore, using comparisons recognizes tacit knowledge of teachers when they are making assessments

(Kimbell 2021). When implemented as a form of peer assessment it might support learning and the transfer of tacit knowledge (Kimbell 2021).

In CJ, the process to scale representations based on the judgments comes from Thurstone's law of comparative judgment (Thurstone 1927), which was Thurstone's attempt to develop a new way for scaling in educational tests (Thurstone 1925). Pollitt and Murray (1995) reintroduced this method to assess the level of language proficiency. Attention for CJ has been rising steadily since and has seen an apparent surge in the last decade. It has, for example, been used for the assessment of academic writing (van Daal et al., 2017), visual arts (Newhouse 2014), graphic design (Bartholomew et al., 2019), and mathematical problem solving (Jones and Alcock 2014). CJ can also be deployed in a peer assessment both as an assessment tool and as a learning tool (e.g., Bouwer et al., 2018; Bartholomew et al., 2019; Brignell et al., 2019).

The method of CJ has also been successfully applied in the context of standard maintaining (e.g., Bramley and Gill 2010; Curcin et al., 2019) and comparability of standards (e.g., Bramley 2007; Holmes et al., 2018) in UK national assessments. For standard maintaining, in order to equate the difficulty of exams over two consecutive years, representations of 1 year are paired with those of the next year. Based on the results of this CJ exercise equal grade boundaries are determined over both years, eliminating differences in difficulty between those years. For comparability of standards, representations from two examining boards are paired in order to investigate if the assessment results are comparable across boards. While the use of CJ for standard maintaining is fairly recent, it was already used for comparability of standards from 1997 (Bramley 2007).

Pollitt and Murray (1995) and Bramley, Bell, and Pollitt (1998) recognized early on that the method of CJ is highly inefficient, needing a lot of comparisons for the results to reach an acceptable reliability level. In CJ the reliability is measured by the scale separation reliability (SSR), reflecting how accurate the ability score estimates are. The SSR can be interpreted as in how far assessors agree with each other regarding the quality of the representations (Verhavert et al., 2018). A meta-analysis showed that in order to reach a scale separation reliability (SSR) of 0.70, at least 14 comparisons per representation are needed. For an SSR of 0.90 this rises to 37 comparisons per representation (Verhavert et al., 2019). In educational practice, this often leads to a large total number of comparisons (N_C) considering the regular sizes of student groups, which impedes the practical feasibility of the method. Bramley et al. (1998) have accurately summarized this problem, as follows: "The most salient difficulty from a practical point of view is the monotony of the task and the time it takes to get a sufficient number of comparisons for reliable results." (14).

Up until 2008 (Kimbell et al., 2009; Kimbell 2021), in CJ, pairs were constructed using a semi-random selection algorithm (SSA). This algorithm prefers representations that have appeared in pairs the least number of times. It pairs those with representations they have not yet been compared with. At the request of Kimbell et al. (2021), Pollitt developed an adaptive selection algorithm (ASA) to construct pairs based on information of previous comparisons as a solution to the efficiency problem. This ASA (Pollitt 2012) is inspired by computerized adaptive testing (CAT; see also below). It pairs representations which have ability scores that are close together. For this, it used preliminary ability scores estimated based on previous comparisons within the assessment and the Fisher's information statistic. In CJ, the Fisher's information is highest when the difference in ability scores between the pairs is lowest. With this adaptive selection algorithm Pollitt claims that very high SSR values of 0.94 or above, can already be obtained after around 13 comparisons per representation (Newhouse 2014). This is not only a large gain in efficiency, but also presents a higher reliability compared with traditional marking.

There are, however, some concerns regarding Pollitt's ASA and the related reliability. Bramley (2015) and Bramley and Vitello (2019) have shown that this type of adaptivity inflates the reliability of the assessment. This is probably because this type of adaptivity "capitalizes on chance" (Bramley and Vitello 2019, 45), Namely, the construction of pairs is based on information that is in itself not reliable because it only consists of a few data points (Bramley and Vitello 2019, i.e., preliminary ability score estimates based on previous judgments within the assessment). In this way, the adaptivity reduces the chance that contradictory information is collected. While the adaptivity may genuinely reduce the standard error (se) of the estimates, it artificially inflates the standard deviation (SD). Since the SSR formula is based on the difference between SD and se (see below), this means the SSR is inflated by an unknown amount and cannot be used on its own as a measure of reliability (Bramley and Vitello 2019; Crompvoets et al., 2020).

The decrease in *se* and increase in SD has some consequences for the use of the assessment results in practice. The *se* reduction means that when we are merely interested in the ranking of the students (and not the estimated value), adaptivity might not pose that much of a problem (Bramley and Vitello 2019)[1]. On the other hand, if we use the estimated ability values of the students as scores in high stakes assessments, the increased SD does pose a problem, namely, it means that the estimated ability values are shifted away from 0 and thus away from their *true value* (Bramley and Vitello 2019). In any case, when using adaptivity we can no longer rely on the SSR value to indicate how reliable, i.e., how certain, the results are.

While in CAT there is no issue of the inflation of the reliability coefficient, this is rather problematic for CJ contexts. This is because of two related, important differences between CAT and CJ. A first difference lies in the differing background of both. CAT is based on Item Response Theory (IRT) where CJ is based on the BTL-model. IRT models compare the difficulty of test items (e.g., exam questions) with the ability of students (to answer these questions). In CJ the performance of students is compared directly. As such, the BTL-model compares student abilities.

[1] However, because of an inflated reliability coefficient the danger does exists that one places more confidence in the results than is warranted. It is always desirable to aim for results that are as close to the true ranking and rank order as possible.

Second, the item difficulty in CAT is determined before any assessments are conducted. This is done in an extensive calibration study with several hundreds of students. As such, if a student is presented with an item that has a difficulty close to the preliminary ability of that that student using, for example, Fisher information, that item difficulty is a fixed value on a pre-calibrated scale. Therefore, the eventual estimate of the ability of the student and the reliability of this estimate are not biased by the adaptivity of the selection algorithm. On the contrary, in Pollitt's ASA for CJ, the abilities of the students in any pair are not yet fixed. As such, Pollitt's algorithm for CJ capitalized on chance where CAT algorithms do not.

The current article investigates a new ASA to address the efficiency problems in CJ (detailed description see following section). In order to overcome the problems with Pollitt's type of adaptivity, the ASA is more strictly based on CAT algorithms. Namely, like CAT has test items with pre-calibrated, fixed difficulty scores, the newly proposed algorithm makes use of a set of representations with pre-calibrated, fixed ability scores. This is further referred to as the reference set. New representations are initially paired with a randomly selected representation in the middle of the reference set. Based on this first comparison, a preliminary ability value is estimated for the new representations. Consecutively, the new representations are individually paired with representations of the reference set with ability scores close to the preliminary scores of the new representations based on the Fisher information statistic. The ability scores of the new representations are consecutively updated and new pairs are constructed as before. This is repeated until the preliminary estimates reach a certain reliability or the representations are compared a predetermined maximum number of times.

The idea of using a fixed reference set was suggested by Bramley (2015) and Bramley and Vitello (2019) as a solution to the SSR inflation in Pollitt's ASA because the fixed ability scores in the reference set anchor the estimates of the new representations. The new ASA might not only have the advantage of countering SSR inflation, it might also provide advantages when using CJ for maintaining standards. Specifically, it would only be necessary to scale representations for 1 year and set grade boundaries. In all consecutive years, this reference set can be used to scale the new representations with fewer pairwise comparisons, provided that the assessment task is similar. In this way, it might even be possible to expand the scale or (gradually) replace representations of previous years with newer ones.

The Reference Set-Based Adaptive Selection Algorithm

The reference set-based adaptive selection algorithm (RSB-ASA) places new representations on a measurement scale consisting of a pre-calibrated set of representations (of the competence under assessment), called the reference set. It does this in two preparatory steps and four actual steps. For simplification, the algorithm will be described from the standpoint of a single, new representation. Mind that, in practice, multiple new representations are assessed. Furthermore, with multiple representations an assessment can theoretically be divided into rounds. One round ends when every active, new representation has gone through all steps, with active meaning involved in comparisons (see also **Step 4B**). For clarity, the rest of the article representations in the reference set will be indicated with the letter j and, when needed, k, and their fixed ability values with α_j and α_k. New representations are indicated with i and their ability value with v_i.

Step A: A reference set with an acceptably high reliability is constructed. A CJ assessment using the SSA for constructing pairs is conducted in order to pre-calibrate the ability scores of the representations in the reference set. As is common, ability scores are estimated using the BTL model (Bradley and Terry 1952; Luce 1959):

$$P_k(\alpha_j) = p(X_{jk} = 1|\alpha_j, \alpha_k) = \frac{exp(\alpha_j - \alpha_k)}{1 + exp(\alpha_j - \alpha_k)} \tag{1}$$

with $P_k(\alpha_j)$ or $p(X_{jk} = 1|\alpha_j, \alpha_k)$ the probability that representation j is preferred over representation k and, thus, $X_{jk} = 1$ meaning that representation j is preferred over representation k, and α_j and α_k the ability scores (in logits) for representation j and representation k, respectively. In the RSB-ASA, these ability estimates are fixed. The reliability of the ability score estimates for the representations in the reference set is measured using the rank-order SSR (Bramley and Vitello 2019)[2]:

$$SSR = \frac{\sigma_\alpha^2 - MSE}{\sigma_\alpha^2} \tag{2}$$

with σ_α the standard deviation of the estimated ability values and MSE the mean squared standard error calculated as:

$$MSE = \frac{\sum_j^n se_{\alpha_j}^2}{n}, \; with \; j = k \tag{3}$$

with se_{α_j} the standard error of estimate, calculated as (Wright and Stone 1999):

$$se_{\alpha_j} = \frac{1}{\sqrt{\sum_{j,\, k \neq j} I_{jk}}} \tag{4}$$

with I_{jk} calculated as

$$I_{jk} = P_k(\alpha_j)(1 - P_k(\alpha_j)) \tag{5}$$

with $P_k(\alpha_j)$, calculated as in **(1)**.

The rank-order SSR of the reference set (further referred to as the reference set SSR) should be high enough. What this means will be investigated in Study 1. The size of the reference set should be as large as needed to have a measurement scale that is fine grained enough. In the current article the authors went for a

[2] The rank-order SSR is commonly just referred to as the SSR. However, in this article we refer to it as the rank-order SSR in order to disambiguate it from the point or estimated SSR used further in this section.

reference set of 140 (Study 1) and 200 (Study 2) representations. However, the optimal size of the reference set goes beyond the scope of the current article.

Step B: A subset of the reference set is determined. This subset consists of representations with an ability score close to 0. An ability score of 0 is the score of an average representation in the reference set and as such the best starting point for comparing a new representation.

From here, the actual algorithm starts. Note that by definition the fixed ability values of the reference set are not re-estimated hereafter.

Step 1: A new representation i is randomly paired with a representation j from the subset determined in **Step B**.

Step 2: A preliminary ability score v_i is estimated for representation i using the BTL model:

$$P_j(v_i) = p\left(X_{ij} = 1 | v_i, \alpha_j\right) = \frac{exp\left(v_i - \alpha_j\right)}{1 + exp\left(v_i - \alpha_j\right)} \tag{6}$$

with $P_j(v_i)$ or $p(X_{ij} = 1|v_i, \alpha_j)$ as the probability that representation i is preferred over representation j, $X_{ij} = 1$ representation i is preferred over representation j, and v_i and α_j are the ability scores (in logits) for the new representation i and the reference set representation j, respectively. Parameter α_j is now fixed, and parameter $\boldsymbol{v_i}$ is used for the variable parameter. Otherwise, the formula is in fact equivalent to formula **(1)**. Index j is equal to every representation that representation i has been compared with. Note that, in practice, this step is executed once every new representation i has been in a pair once.[3]

Step 3: Is the predetermined value of the stopping criterion reached or exceeded for representation i? There are two types of stopping criteria, fixed criteria and variable criteria. With fixed criteria, all representations are compared an equal number of times. This comes down to setting a fixed number of comparisons per representation (N_{CR}). With variable criteria, each representation is compared a different number of times. In the current algorithm, the accuracy of the preliminary ability estimate of representation i is used[4]. In CAT algorithms, generally, the standard error of estimation (*se*) is used as a measure of estimate accuracy. The current algorithm resorts to the reliability of the ability estimate of representation i. In the BTL-model, this gives equivalent results to the *se*, but is easier for practitioners to interpret. In order to measure the reliability of the ability estimate (of a single representation), the point SSR_i or the estimated SSR_i is calculated:

$$SSR_i = \frac{\sigma_\alpha^2 - se_i^2}{\sigma_\alpha^2} \tag{7}$$

with σ_α the standard deviation of the fixed ability values (in the reference set) and se_i the standard error of estimate of representation i. As can be noted, the above formula for the estimated SSR_i differs from that of the rank-order SSR [formula **(2)**] in that the MSE has been replaced by the *se* of the ability estimate v_i of representation i.

Returning to the question posed in **Step 3**: is the predetermined value of the stopping criterion reached or exceeded for representation i? If not, continue to **Step 4**. If yes, stop here. This representation no longer appears in pairs in this assessment. For fixed stopping criteria, this happens for all representations at once. For variable stopping criteria, this is determined for each representation separately. It is, however, possible that some representations never reach the stopping criteria. Therefore, with a variable stopping criterion, a maximal N_{CR} must be set to prevent the algorithm from continuing forever.

Step 4: Select representation j providing the most information for representation i is. Information is measured here with the Fisher information criterion.

Step 4A: The Fisher information I_{ij} is calculated for all representations j in the reference set, against the ability of representation i (Wright and Stone 1999):

$$I_{ij} = P_j(v_i)\left(1 - P_j(v_i)\right) \tag{8}$$

with $P_j(v_i)$, calculated as in **(6)**, the predicted probability that representation i will be preferred over representation j given the ability scores v_i and α_j. This formula is equivalent to formula **(5)**.

Step 4B. Representation i is paired with the representation j that has the largest value for I_{ij}. With the BTL-model this generally comes down to the representation i with an ability score closest to the ability score of representation j.

The Current Research

The current research investigates the efficiency and accuracy of the RSB-ASA described in the previous section. Specifically, it attempts to answer the following research questions (RQ):

RQ1 Does using the RSB-ASA in a CJ assessment 1) lead to a higher efficiency, 2) while producing results with the same accuracy as the SSA?

RQ2 Does the RSB-ASA produce an inflation in standard deviation of the CJ results?

RQ3a Does the reference set reliability in the RSB-ASA influence the efficiency and the accuracy of the results?

RQ3b Does the reference set reliability in the RSB-ASA influence the standard deviation of the CJ results?

[3] Notwithstanding that the (preliminary) ability values of the new representations i are estimated for all representations at once, it is possible to estimate the ability value of every representation i separately. This is possible because the ability values of the representations j in the reference set are fixed.

[4] Other variable stopping criteria are possible, like estimate stability and information gain (the difference between the maximum information value for this representation in this round and the maximum information value for this representation in the previous round). These other stopping criteria are topics for further research.

RQ4a Does the stopping criterion, fixed or variable, in the RSB-ASA influence the efficiency and the accuracy of the CJ results?

RQ4b Does the stopping criterion, fixed or variable, in the RSB-ASA influence the standard deviation of the CJ results?

This was done in two studies. In Study 1, assessors created a reference set in a CJ study using the SSA. From this reference set, a subset of representations was selected to be placed back on the reference set in the second part of this study. This provided answers to RQ1 and RQ2. Furthermore, in Study 1, the reliability of the reference set was manipulated (RQ3) and different stopping criteria were used (RQ4) when analyzing the data. For more details, see the *Methods* section of Study 1.

Because Study 1 involved real assessors (i.e., was not a simulation), it was not feasible to include replications, due to practical constraints. The reference set was also derived from an assessment with the same assessors who later conducted the assessment implementing the RSB-ASA. Last, the fixed stopping criterion and the maximal number of comparisons with the variable stopping criteria set in Study 1 might have been too restrictive. In order to address these shortcomings (discussed in some more detail in the *Discussion* section of study 1), a simulation study was conducted as Study 2.

Study 2 looked into the efficiency of the RSB-ASA and the accuracy of the results (RQ1) by comparing the results of the simulation with the RSB-ASA with those of a simulation with the SSA. Also, the standard deviation of the results of both simulations was calculated and compared with each other and with the standard deviation of the generating values (RQ2). Also here, different stopping criteria were used (RQ4). For more details, see the *Methods* section of Study 2.

In both studies, efficiency was conceptualized as the N_{CR} were needed in the CJ assessment. Accuracy was conceptualized as the average difference between the resulting ability estimates of the new representations with the RSB-ASA and the, so called, true ability scores. For details, see the section on the measures in the *Methods* sections of each study. It is expected that reference sets with a higher reliability will result in a higher efficiency and a higher accuracy of the estimates. Using a predetermined N_{CR} as the stopping criterion should also result in a higher accuracy of the estimates. However, it might lead to a reduction of the efficiency compared with a variable stopping criterion. It will be important here to see if the gain in accuracy weighs up against the decrease in efficiency.

STUDY 1

Method

Materials

As representations, 160 short essays were selected from a total of 7,557 essays. This number was chosen to keep the work for assessors doable and keep the paid work hours within budget. Furthermore, based on the experience of the authors, this number should lead to a reference set that is fine grained enough for use in the algorithm. What the optimal size of the reference set should be, goes beyond the current article. The 160 representations were selected at random by means of the select cases tool in SPSS (IBM Corp. 2016). The essays were taken from the Economics Higher Level Paper 1 (time zone 2[5]) exam of May 2016. Specifically, they were all responses to question 1b (Q1b): "Evaluate the view that regulations are the most effective government response to the market failure of negative externalities." This was a subpart of an optional question and was worth 15 out of a total of 50 marks. There was no word or time limit for the essay, although the total exam time was 90 min.

All pages not containing a response to Q1b were removed and when the response to this question began or finished part way through a page, any writing relating to other questions on the exam was covered from view. The essays were then anonymized and all examiner marks and comments were removed.

Participants

The reference set was created in a CJ assessment (algorithm **Step A**) including 15 assessors. From these 15 assessors, only 10 were available to participate in the assessments implementing the RSB-ASA. These sample sizes were chosen in order to keep the workload for assessors manageable and at the same time have a decent proportion of judgments attributed to each judge, making sure that the judgments of each assessor had a realistic weight in the end result. The latter is also the reason to reduce the number of assessors from 15 to 10, as the N_{CR} was less. These decisions are, however, based on the experience of the authors. To our knowledge, research regarding the effect of the number of assessors on the final results of a CJ assessment are currently lacking.

The 15 assessors were all existing IB examiners, and were recruited by e-mail. Of the 15 assessors, 12 had marked the question during the May 2016 examination session and the remaining three either marked another question on this examination, or marked questions from the time zone 1[5] variant of the examination. The 15 assessors also included the Principal Examiner for the examination, who is responsible for setting the overall marking standard, and two "team leaders" who are considered reliable and experienced examiners and have responsibility within an examination session for leading a small team of examiners.

All assessors were paid for their work, and all signed a "Contributor's Agreement," which included permission to use their anonymized judging data.

Procedure

There were two phases in this study: In phase 1, the reference sets were constructed (cfr. algorithm **Step A**), and in phase 2, new representations were compared with the reference sets using the RSB-ASA as described above. Phase 1: The assessment used to construct the reference set (algorithm **Step A**) took place in 2017. It was planned to collect 30 judgments per representation,

[5] In some subjects, the IB produces two different versions of an exam, with different variants going to different countries (with different time zones) in order to mitigate academic honesty risks.

TABLE 1 | For each reference set (SSR) the standard deviation of ability estimates (α, $n = 140$), mean and standard deviation of standard error of estimate (*se*, $n = 140$) and the number of comparisons per representation to reach this reference set (N_{CR}).

SSR	α	*se*	N_{CR}
	SD	M (SD)	
0.50	2.75	1.59 (1.07)	8
0.70	2.64	1.25 (0.68)	10
0.80	2.33	0.95 (0.42)	13
0.91	2.13	0.56 (0.17)	30

Note. SSR = rank-order reliability; SD = standard deviation; M = mean.

totaling in 2,400 judgments. Each assessor was therefore asked to make 160 judgments each. Pairs were constructed using the SSA selecting representations randomly, preferring those with the least N_{CR} that have not yet been compared with each other. Assessors were given 3 weeks, between the beginning of April 2017 and the beginning of May 2017 to complete all the judgments. Because the assessors were distributed across the world and many were fitting the work around other commitments, no constraints were placed upon when the assessors were to do the judging within that 3-week window. However, the assessors were asked to attempt to make their judgments as much as possible at the same time (i.e., on set dates and times).

The first phase, the assessment with the SSA, resulted in a rank order with a rank-order SSR of 0.91 (rank-order SD = 2.12; mean parameter $se = 0.56$; $n = 160$; for detailed results, see **Supplementary Table S1** in Supplementary Materials). Twenty representations were taken out of this rank order to be placed back on the reference set the assessments using the RSB-ASA, leaving 140 representations to construct the reference sets. It was opted to have an even spread of representation along the rank order. Therefore, representations at fixed ranks (4th, 12th, 20th, etc.) were selected. In this way the average logit distance between the selected representations was 0.44 (min = 0.17; max = 1.38).

In order to look into the effect of the reference set reliability in the RSB-ASA, it was decided to construct four reference sets with reference set SSR of 0.50, 0.70, 0.80, and 0.91, respectively. The reference sets were based on the judgement data from phase 1. For the three first reference sets, it was determined after how many comparisons (with each representation having an equal number of comparisons) the ability estimates reached a rank-order SSR values of 0.50, 0.70, and 0.80. At each of these rank-order SSR values, the corresponding ability estimates were recorded for all 140 representations. The fourth reference set were the estimates of the 140 representations at the end of the assessment with the SSA. Rank-order SD, mean parameter *se* and N_{CR} are presented in **Table 1** (for detailed results, see **Supplementary Table S1** in Supplementary Materials).

Because there were no constraints placed on when assessors made their judgments, there was no equal distribution of the judgments of the assessors throughout the assessment. Some judgments of the assessors were clustered at the beginning of the assessment, while judgments of other assessors were clustered at the end and still others had an equal distribution throughout the assessment. Therefore, to make sure that each reference set is approximately based on an equal amount of judgments of every assessor, some reordering of the dataset had to take place before the three reference sets with smaller reference set SSRs could be determined. The reordering went as follows: The judgments were first ordered chronologically on time of completion and were then divided into groups of 80 comparisons. Finally, these groups were sorted in a random order. For a more detailed representation of how the judgments of the assessor were distributed throughout phase 1, see **Figures 1**, **2**.

Phase 2: Four assessment sessions were organized implementing the RSB-ASA described above, one for each reference set. As each reference set consisted of the same representations and the representations to be placed back were the same across the sessions, the only difference for assessors was the pairings of the representations. However, in order to make sure that the results in any session would not be influenced too much by the judgments of a few assessors, the order of the sessions was not counterbalanced. The number of judgments was fixed on 10 judgements per representation in each session. This resulted in a total of 200 judgments per session or 20 judgments per assessor per session. The judges got 4 weeks to complete their judgments (between mid-August and mid-September 2017). **Figure 3** presents how the judgments of the assessors were distributed throughout phase 2 in more detail.

All assessments were conducted in and controlled by the D-PAC[6] platform. The assessments were conducted under the supervision of AF.

Afterward, the judgment data was processed as follows. For every assessment session implementing the RSB-ASA three stopping criteria were implemented: 10 comparisons per representation (further called fixed stopping criterion) and an estimate reliability of 0.70 and 0.90 (further called, respectively, estimated SSR_i 0.70 and estimated SSR_i 0.90). These reliability levels were used because these are the reliability levels commonly aimed for in formative and summative assessments, respectively (Nunnally 1978; Jonsson and Svingby 2007). Thus, for all 20 representations, ability scores were estimated after 10 judgments per representation. For estimated SSR_i 0.70 and estimated SSR_i 0.90, the method was as follows. After every round[7] the SSR_i is calculated [as in formula **(7)**; cfr. **Step 3**] for the preliminary ability value of each of the 20 representations (cfr. **Step 2**). If this SSR_i equals or exceeds 0.70 or 0.90, respectively, the corresponding ability value was noted, as well as the N_{CR} needed to obtain this value. If a representation did not reach an estimated SSR_i of 0.70 or 0.90 after 10 comparisons, the ability value after 10 comparisons per representation was recorded. Ability scores were estimated using a joint maximum likelihood algorithm with an epsilon bias correction factor of 0.003 (for details, see Verhavert 2018).

[6] Currently named Comproved; https://comproved.com/en/.

[7] A round is defined as the moment where every representation has been in a pair an equal amount of times. With 20 representations this is after 10 comparisons, 20, 30, . . .

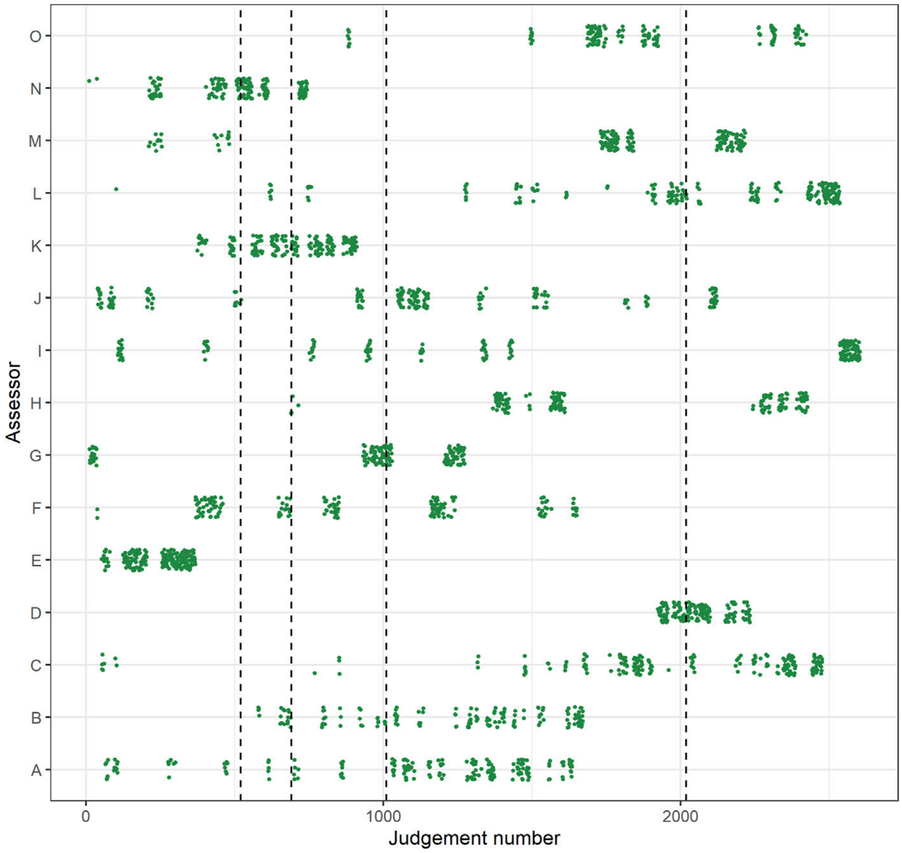


FIGURE 1 | Distribution of judgments throughout the assessment in phase 1 (Judgment number; x-axis) by assessor (y-axis), before reordering. The vertical dotted lines indicate the judgment number at which the estimates reached a rank-order reliability of 0.50, 0.70, 0.80, and 90 (respectively from left to right). Note: jitter added to y-axis coordinates for clarity.

Measures

For the results of every reference set and all stopping criteria, the accuracy of the noted ability scores (of the 20 representations that were placed back) was calculated by the root mean squared error (RMSE). This is the mean difference between the estimated ability score and the ability score obtained at the end of phase 1. Furthermore, standard deviation (SD) of the ability estimates in every condition in phase 2 was calculated and compared with the SD of the 20 representations at the end of phase 1. This is to see whether there might be a SD inflation in the estimates of the representations that are placed back on the scale, as is the case with ASA's that do not use a reference set.

All analyses were conducted in R (R Core Team, 2020).

Results

From the plot showing the RMSE (**Figure 4**) it is clear that, in general, the RMSE values are the lowest when the stopping SSR_i is 0.90 or when a fixed stopping criterion of 10 comparisons per representation is used. This indicates that the stopping SSR_i 0.90 and the fixed stopping criterion lead to more accurate results compared with the stopping SSR_i 0.70. A second observation is that the RMSE is smaller when the reference set SSR is larger, showing that a more reliable reference set leads to more accurate results. Furthermore, there appears no or just a small difference in RSME between reference set SSR's 0.80 and 0.91. This observation could be explained by assessor fatigue, namely, because the assessment with the 0.91 reference set was presented last, it is possible that the assessors made more mistakes due to fatigue causing a minor drop the accuracy of the results (a higher RMSE) rather than the expected rise in accuracy (a lower RMSE). This interpretation could be verified by calculating assessor misfit, a measure for the number of mistakes an assessor makes weighted by the severity of the mistakes[8]. However, because of the assessment setup and the RSB-ASA, there is no longer an equal distribution of the difficulty of the comparisons[9] between assessors. Therefore, the misfit values are no longer comparable. As a third observation, the difference in RMSE between either the fixed stopping criterion or the stopping SSR_i of 0.90 and the stopping SSR_i of 0.70 decreases when the reference set SSR becomes larger. However, these results might be

[8] The further apart representations in a pair are with regard to estimated ability score, the more severe the mistake.

[9] A comparison is more difficult when the representations in the pair lie closer together with regard to estimated ability score.

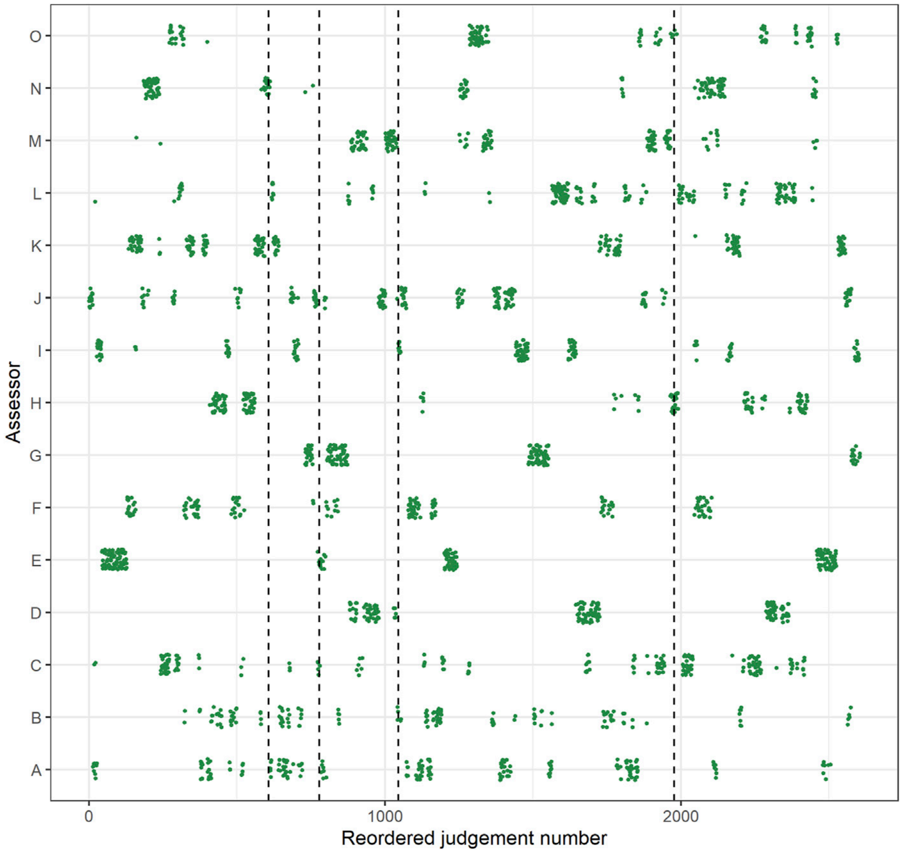


FIGURE 2 | Distribution of judgments throughout the assessment in phase 1 (Judgment number; x-axis) by assessor (y-axis), after reordering. The vertical dotted lines indicate the judgement number at which the estimates reached a rank-order reliability of 0.50, 0.70, 0.80, and 90 (respectively from left to right). Note: jitter added to y-axis coordinates for clarity.

an artifact of the assessment setup, namely, as the results discussed in the next paragraph show, the limit of 10 comparisons per representation might be too low in order for the representations to reach a stopping SSR_i of 0.90.

As expected, it was observed that fewer comparisons per representation are needed when estimated SSR_i was used as stopping criterion than with the fixed stopping criterion (**Figure 5**). This is true for both the median and the mean N_{CR} (respectively, the filled diamond and the square in **Figure 5**). Specifically, fewer comparisons per representation were needed to reach the estimated SSR_i of 0.70, making this the most efficient stopping criterion. For the estimated SSR_i of 0.90, it is observed that, overall, 11 of the 20 representations reach this SSR_i level before 10 comparisons per representation, the rest never reached this value. Moreover, at the fixed stopping criterion the average estimated SSR_i was 0.88 (SD = 0.02), which is lower than the stopping criterion SSR_i of 0.90. This shows that 10 comparisons per representation might have been a bit too low for a fixed stopping criterion.

Taking the total number of comparisons into account (**Table 2**), less comparisons are needed for the stopping criterion SSR_i of 0.70 compared with the stopping criterion SSR_i of 0.90 and the fixed stopping criterion, which was to be expected. This shows that the stopping criterion SSR_i of 0.70 is the most efficient. Moreover, the total number of comparisons increases as the reference set SSR increases. Thus, a more accurate reference set appears to reduce the efficiency of the algorithm. There are several possible explanations (for details, see the *Discussion* section). Additionally, with a reference set SSR of 0.50, there still is a small difference in total number of comparisons between stopping criterion SSR_i of 0.90 and the fixed stopping criterion. This difference decreases as the reference set SSR increases again pointing in the direction that 10 comparisons per representation might have been too restrictive. Finally, because of the difference between the estimated SSR_i and the rank-order SSR, the N_{CR} in **Table 2** cannot be compared with those needed to reach a rank-order SSR of 0.70 and 0.90 in phase 1 (**Table 1**). This should be further looked into in study 2.

As a final observation, the SD of the ability estimates, of the 20 representations that were placed back on the reference set, becomes smaller as the reference set SSR becomes larger (**Table 3**). Contrary to our expectations, there is only a minimal, negligible difference in SD between the stopping criteria within every reference set. When the reference set SSR is 0.90, the SD of the ability estimates approaches the SD of the

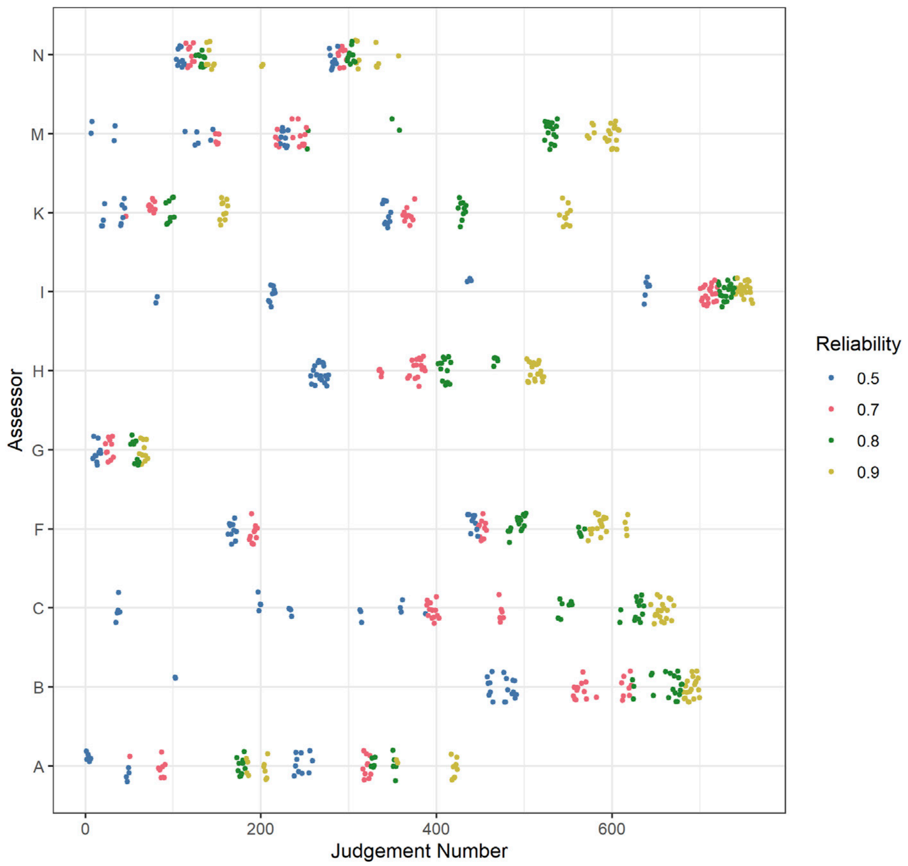


FIGURE 3 | Distribution of judgments throughout the assessments in phase 2 (Judgment number; x-axis) by assessor (y-axis), per reference set reliability (colors). Note: jitter added to y-axis coordinates for clarity.

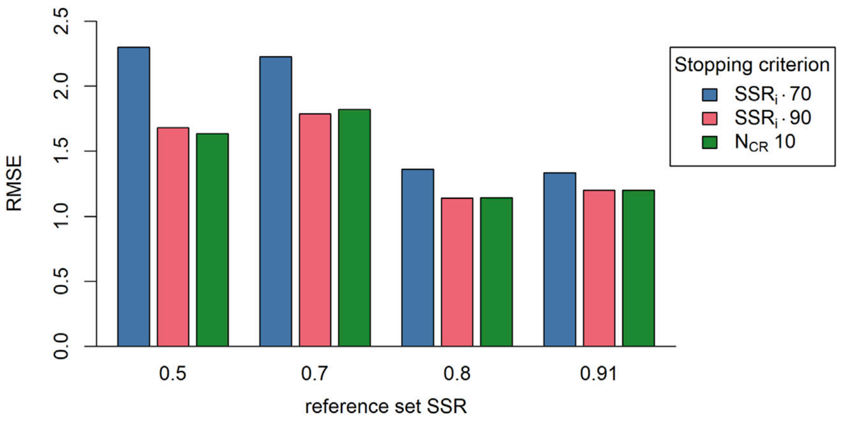


FIGURE 4 | The root mean squared error (RMSE; *n* = 20) per reference set (SSR of reference sets) and per stopping criterion (colors; SSR$_i$ 0.70, SSR$_i$ 0.90, N$_{CR}$ 10). Note: SSR$_{i,}$ = estimate reliability; N$_{CR,}$ = number of comparisons per representation.

ability values of the selected representations obtained in phase 1, namely, 2.15. With a low reference set SSR, it can be said there is an increase in SD [Δ_{SD} = (0.86; 1.11)]. However, with a high reference set SSR the difference in SD is almost negligible [Δ_{SD} = (0.04; 0.12)].

Discussion

The abovementioned results tentatively show that the RSB-ASA is more efficient than the SSA used in phase 1. The largest efficiency gain can be made by using an estimated SSR$_i$ of 0.70 as a stopping criterion. There appears however a tradeoff between

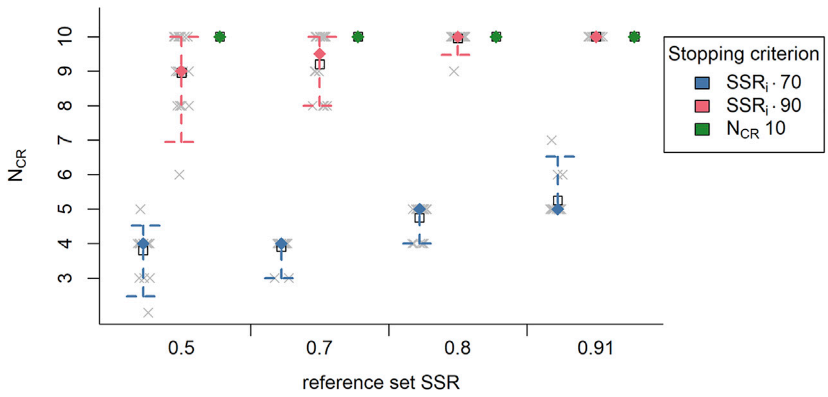


FIGURE 5 | The median (*n* = 20; filled diamond) and average (*n* = 20; empty square) number of comparisons per representation (Ncr) per reference set (SSR of reference sets) and per stopping criterion (colors; SSR_i 0.70, SSR_i 0.90, N_{CR} 10) with the interval covering 95% of values (*n* = 20; dashed bars) and the actual number of comparisons per representation (grey x). Note: SSR_i = estimate reliability; $N_{CR,}$ = number of comparisons per representation.

TABLE 2 | Total number of comparisons per reference set (SSR) and per stopping criterion.

	Stop criterion		
Reference set SSR	SSR_i 0.70	SSR_i 0.90	N_{CR} 10
0.50	76	179	200
0.70	78	184	200
0.80	95	199	200
0.91	105	200	200

Note. SSR = rank-order reliability; SSR_i = estimated reliability; N_{CR} = number of comparisons per representation.

TABLE 3 | The standard deviation (SD; *n* = 20) of the ability estimates of the selected representations per reference set SSR and stopping criterion.

	Stop criterion		
Reference set SSR	SSR_i 0.70	SSR_i 0.90	N_{CR} 10
0.50	3.26	3.03	3.40
0.70	3.40	3.23	3.26
0.80	2.92	2.88	2.88
0.91	2.27	2.23	2.23

Note. SSR = rank-order reliability; SSR_i = estimated reliability; N_{CR} = number of comparisons per representation.

the efficiency and the accuracy. The stopping criterion estimated SSR_i of 0.70 is less accurate than using an estimated SSR_i of 0.90 as a stopping criterion or even a fixed stopping criterion of 10 comparisons per representation. There is also an effect of the accuracy of the reference set used in the adaptive algorithm. The higher the accuracy of the reference set, expressed by the reference set SSR, the higher the accuracy of the assessment results. The latter will be discussed in further detail in the general discussion.

It was also observed that a more accurate reference set led to an increase in comparisons needed to reach these accurate results. One explanation is that when the reference set values are too accurate, toward the end of the assessment, the assessors receive pairs of representations that are very difficult to distinguish. As a consequence, they might make more judgment errors. Alternatively, it can be assessor fatigue. Because the assessors receive the assessments in the same order (of reference set SSR), they might be more tired with the last assessment, thus, making more errors. In both cases, more errors mean that representations might need more comparisons in order to reach an accurate enough estimate.

It must further be remarked that this study did not contain any replications with the same or a different assessor group. Therefore, it is unsure if the above described differences are due to random error. Second, the assessment conducted to construct the reference sets (phase one in procedure) and the assessments conducted with the RSB-ASA (phase 2 in procedure) were all done by the same assessors (or a subset thereof). Therefore, the assessors might already have been familiar with the representations and the CJ method, which might have influenced the results. Third, results showed that 10 comparisons per representation might have been a bit too strict for a stopping criterion. Finally, due to the incomparability of the rank-order SSR and the estimated SSR_i, it cannot be confirmed if the RSB-ASA is in fact more efficient that the SSA.

STUDY 2

In order to address the shortcomings discussed in the previous paragraph and confirm the findings of Study 1, and to look into the theoretical accuracy and efficiency of the RSB-ASA, a simulation study was conducted in Study 2. This allows to make a large number of (theoretical) replications under highly controlled setting, thereby reducing random errors. Assessments also start from generating values, eliminating the need to construct reference sets and making it possible to compare estimates with true values. Furthermore, Study 2 will use a higher number of comparisons per representation and it will allow a more direct comparison between the RSB-ASA and the SSA.

Method

Generating Values

For the current study, 220 ability values (i.e., the generating values) were randomly sampled. This number was chosen in order to have a reference set that is fine grained enough to use in the algorithm. Again, further research will be needed in order to look into the effect of the size of the reference set. The generating values were sampled from a normal distribution, using the norm function from the stats package in R (R Core Team 2020), with a mean of 0 and SD of 2.12, which was equal to the rank order from Study 1. In general, because of restrictions in the estimation procedures (Bradley, 1976; Molenaar, 1995; Verhavert 2018), CJ assessments result in ability estimates that average to 0. Therefore, the sampled ability values were transformed to average 0. The resulting SD was 2.13. From these generating values, 20 ability values were selected as the theoretical abilities of so-called new representations. First, we selected two representations that have a high probability of winning or losing all comparisons. For this, the third highest ability value and the third lowest ability value were selected. The remaining 18 ability values were selected so that the distance in logits between consecutive new representations is approximately equal. The average distance between consecutive ability values was 0.53 (SD = 0.06). The selected ability values had an SD of 3.15. All ability values that were not selected were used as the reference set. Because the abilities were sampled, it was not possible to use reference sets of different SSR levels. The generating values can be found in **Supplementary Table S2** in the additional materials.

Simulation Study

Two CJ assessments were simulated. In the first, the RSB-ASA was used to construct the pairs. In the second, pairs were constructed using the SSA. As a reminder, this algorithm prefers representations with the least N_{CR} that have not yet been compared with each other. The second CJ assessment was simulated in order to compare the results of an assessment implementing the RSB-ASA with the results of an assessment with a random selection algorithm, which is considered as a benchmark for CJ assessments.

In the assessment using the RSB-ASA, the same three stopping criteria were used as in study 1, namely the fixed stopping criterion, an estimated SSR_i of 0.70 and an estimated SSR_i of 0.90. However, the fixed stopping criterion was increased to 20 comparisons per representation. The preliminary estimate of the ability scores and the N_{CR} of the selected representations were recorded per stopping criterion.

In the assessment using the SSA, it is less straightforward to implement the same stopping criteria as with the RSB-ASA because of two reasons. First, it has been shown that CJ assessments with the SSA only reach a reliability of 0.90 after around 37 comparisons per representation (Verhavert et al., 2019). Stopping after 20 comparisons per representation will lead to unreliable results. Second, calculating the estimated SSR_i is not common for the SSA. Normally, the reliability is calculated over all representations, using the rank-order SSR as in formulas (**6**, **7**). This might, however, give a biased result in comparison with the RSB-ASA. Therefore, two sets of stopping criteria were used. The first set served to increase the comparability of the results with those of the SSA. This set thus consisted of the estimated SSR_i of 0.70, estimated SSR_i of 0.90, and N_{CR} of 20. The second set reflected more common practice using the SSA. It consisted of the rank-order SSR of 0.70, rank-order SSR of 0.90 and N_{CR} of 37. The latter stopping criterion resulted in 4,070 comparisons in total. Preliminary (or intermediate) estimates of the ability scores and the N_{CR} of the selected representations were recorded for every stopping criterion in both sets.

Both simulations were repeated 1,000 times. The simulation was conducted in and controlled by R (R Core Team, 2020).

Measures

As a measure of accuracy, the RMSE of the estimates of the 20 selected representations was calculated against the generating values in every replication for every stopping criterion in the RSB-ASA and the SSA. Ability scores were estimated using a joint maximum likelihood algorithm with an epsilon bias correction of 0.003 (for details, see Verhavert 2018). The RMSEs are discussed in the *Results* section and are presented in the figures, which are averaged over the replications. To measure the efficiency of the RSB-ASA and the SSA, for every replication, the median N_{CR} was registered when either the estimated SSR_i of 0.70 and 0.90 or the rank-order SSR of 0.70 and 0.90 was reached. Also here, the N_{CR} discussed in the results were average over replications. Furthermore, to check for a possible inflation of the SDs, the SD of the ability estimates of the selected representations were calculated per stopping criterion in both the RSB-ASA and the SSA in every replication. Also, the discussed SD values are averaged over replications.

Results

As in Study 1, it is observed that the RMSE is the lowest for the fixed stopping criterion (here, 20 comparisons per representation) and the estimated SSR_i of 0.90. Study 2 shows that this is independent of selection algorithm (**Figure 6A**). Contrary to Study 1, within the RSB-ASA, the fixed stopping criterion led to a lower RMSE than the stopping criterion estimated SSR_i of 0.90, showing that a fixed stopping criterion produces the most accurate results. However, the difference in RMSE is merely 0.11. Within the SSA, both stopping criteria have equal RMSEs. Additionally, the RSB-ASA results in a lower RMSE than the SSA (average difference = 0.24; **Figure 6A**). This is, however, only the case when the stopping criteria of the RSB-ASA are used (i.e., an estimated SSR_i or 20 comparisons per representation). When a rank-order SSR or a fixed stopping criterion of 37 comparisons per representation is used (more common with the SSA), there is no difference in RMSE between the RSB-ASA (**Figure 6A**) and the SSA (**Figure 6B**). Comparing the left set of bars in **Figure 6A** with the set of bars in **Figure 6B** shows that the apparent differences between the fixed stopping criteria (N_{CR} 20 and N_{CR} 37) or between the rank-order SSR and the estimated SSR_i are not significant; the average value falls within each other's 95% CI.

Figure 7A shows that there is no difference in SD of the ability estimates between stopping criteria within algorithms. The SSA does, however, lead to a slightly lower SD overall than the RSB-

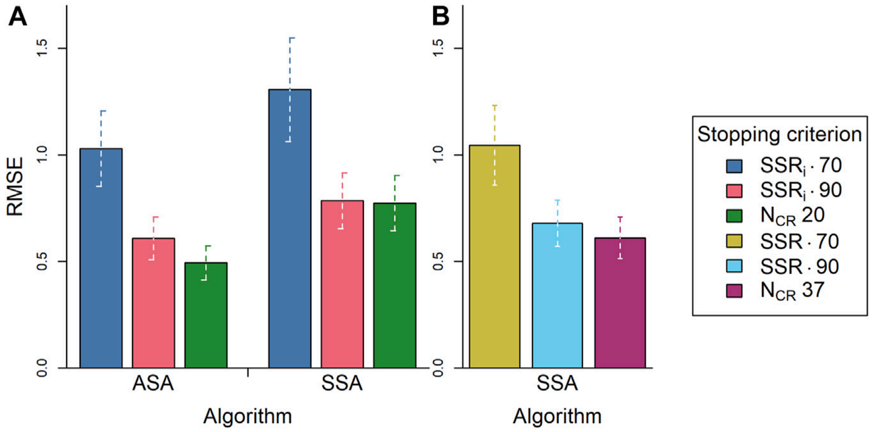


FIGURE 6 | The root mean squared error (RMSE; *n* = 20) averaged over simulation repetitions (*n* = 1,000) per algorithm (ASA, SSA) and per stopping criterion [colors; **(A)** SSR_i 0.70, SSR_i 0.90, N_{CR} 20; **(B)** SSR 0.70, SSR 0.90, N_{CR} 37] with 95% confidence intervals (*n* = 1,000). Note: Panel B contains only SSA data. ASA = adaptive selection algorithm; SSA = semi-random selection algorithm; SSR_i = estimated reliability; N_{CR} = number of comparisons per representation; SSR = rank-order reliability.

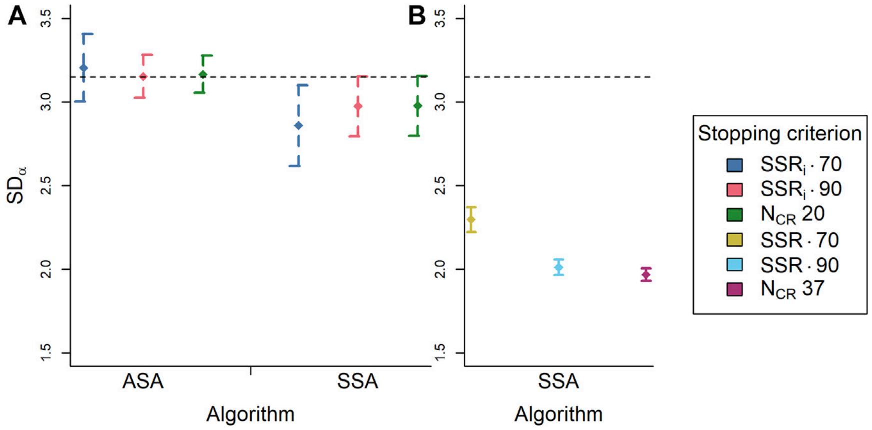


FIGURE 7 | The standard deviation of the ability estimates (*n* = 20) averaged over simulation repetitions (*n* = 1,000) per algorithm (ASA, SSA) and per stopping criterion [colors; **(A)** SSR_i 0.70, SSR_i 0.90, N_{CR} 20; **(B)** SSR 0.70, SSR 0.90, N_{CR} 37] with 95% confidence intervals (*n* = 1,000) and the standard deviation of the generating ability values of the selected representations (*n* = 20). Note: Panel **(B)** contains only SSA data. ASA = adaptive selection algorithm; SSA = semi-random selection algorithm; SSR_i = estimated reliability; N_{CR} = number of comparisons per representation; SSR = rank-order reliability; dashed line = SD of generating values to be placed back (*n* = 20).

ASA. There is no difference between this SD and the SD of the generating values of the selected representations (3.15 as mentioned above; dashed line in **Figure** 7) except for the stopping criterion estimated SSR_i of 0.70 with the SSA, which is lower. With more common stopping criteria for the SSA (rank-order SSR of 0.70 or 0.90 or 37 comparisons per representation; **Figure 7B**) the SDs of the ability estimates are much smaller. The stopping criterion rank-order SSR of 0.70 with the SSA leads to an SD of ability estimates that is a bit larger than the other two common stopping criteria. This SD is still not as large as with the estimated SSR_i or N_{CR} is 10 stopping criteria. It thus appears that the stopping criteria more common for the SSA, cause the results to shift toward the mean.

Figures 8A, B show that with both the SSA and the RSB-ASA the stopping criteria estimated SSR_i and rank-order SSR of 0.70 need the least N_{CR} on average. With the RSB-ASA, it appears that the stopping criterion estimated SSR_i of 0.90 on average needs seven comparisons less than the stopping criterion N_{CR} is 20. Besides, the stopping criterion estimated SSR_i of 0.90 needs more than 10 comparisons per representation, showing that the fixed stopping criterion in Study 1 was too low. Furthermore, with the SSA, the stopping criterion estimated SSR_i of 0.90 needs 20 comparisons per representation, which is higher than with the RSB-ASA. This again shows that the RSB-ASA is more efficient. In addition, looking at common stopping criteria for the SSA (rank-order SSR of 0.70 or 0.90 or N_{CR} is 37), the SSA needs more

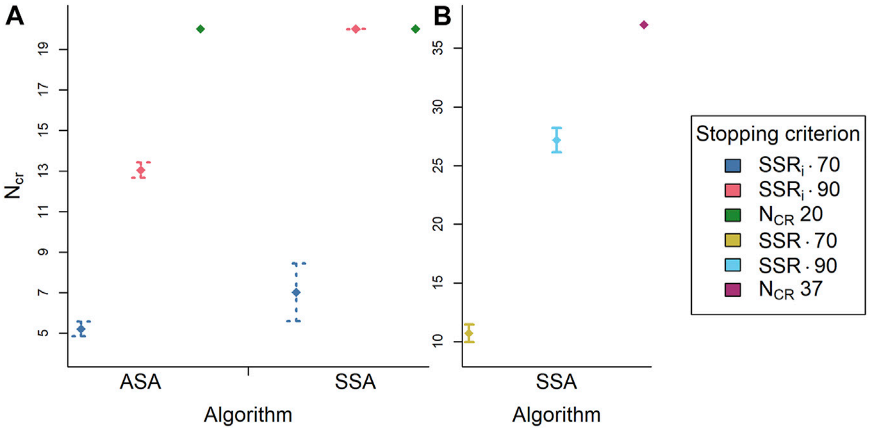


FIGURE 8 | The median number of comparisons per representation (N_{CR}; *n* = 20) averaged over simulation repetitions (*n* = 1,000) per algorithm (ASA, SSA) and per stopping criterion [colors; **(A)** SSR_i 0.70, SSR_i 0.90, N_{CR} 20; **(B)** SSR 0.70, SSR 0.90, N_{CR} 37] with 95% confidence intervals (*n* = 1,000). Note: Panel **(B)** contains only SSA data. ASA = adaptive selection algorithm; SSA = semi-random selection algorithm; SSR_i = estimated reliability; N_{CR} = number of comparisons per representation; SSR = rank-order reliability.

N_{CR} compared with when the RSB-ASA is used. However, because of the incomparability of the estimated SSR_i and the rank-order SSR, no firm conclusions can be drawn from this comparison. In the SSA, there is a difference in N_{CR} between the stopping criteria rank-order SSR of 0.90 and N_{CR} of 37. This shows that the fixed stopping criterion common for the SSA could have been lower.

It was further observed that at 20 comparisons per representation the RSB-ASA reaches an average estimated SSR_i of 0.93 (SD = 0.005). In comparison, the SSA reached an average estimated SSR_i of 0.77 (SD = 0.05) at 20 comparisons per representation. This shows that the RSB-ASA results are more reliable. In addition, at 37 comparisons per representation, the SSA reached an average estimated SSR_i of 0.88 (SD = 0.17), which is also lower than the average estimated SSR_i with the RSB-ASA at 20 comparisons per representation. When the SSA reached a rank-order SSR of 0.70 and of 0.90, the estimated SSR_i averaged 0.52 (SD = 0.13) and 0.84 (SD = 0.03), respectively. Based on the estimated SSR_i, the estimates are less reliable compared with what we would expect based on the rank-order SSR. This seems contradictory because one would intuitively expect that the average estimated SSR_i should approach or approximate the rank-order SSR. However, these two reliability measures are not directly comparable. This will be further elaborated in the discussion.

Discussion

The results of Study 2 confirm that the RSB-ASA produces more accurate results than the SSA when the stopping criteria for the RSB-ASA are used with both algorithms, even with a fixed stopping criterion of 20 comparisons per representation, which was higher than in Study 1. However, the stopping criteria for the RSB-ASA (i.e., estimated SSR_i and 20 comparisons per representation) are not common practice for the SSA. When the stopping criteria common for the SSA (i.e., rank-order SSR and 37 comparisons per representation) were used, the results with the SSA were as accurate as those with the RSB-ASA. Furthermore, the RSB-ASA is also shown to be more efficient than the SSA, independent of the stopping criteria used, confirming the results of Study 1. Also, the stopping criterion estimated SSR_i of 0.70 is the most efficient, independent of algorithm. In combination with the results on the accuracy of the estimates, this illustrates an efficiency-accuracy tradeoff. It needs to be kept in mind that these results leave aside the initial effort of calibrating the reference set.

It was also checked if the RSB-ASA causes an inflation of the SD, which might influence the usability of the estimated ability values. Although the results with the RSB-ASA are as accurate as those with the SSA when common stopping criteria are used for both algorithms, the RSB-ASA results in a higher spread of the estimates (as shown by a higher SD) compared with the SSA. When the results are compared with the SD of the generating values the RSB-ASA does not produce a higher spread of the results. An explanation for these apparently conflicting observations might be that the simulation study uses a perfect assessor. In other words, one whose judgments are exactly in accordance with the Bradley–Terry model. With the SSA and more comparisons per representation, this might cause the estimates to shift toward the mean. This effect should be further investigated.

A third large observation is that, dependent on the reliability measure (i.e., estimated SSR_i and rank-order SSR), the estimates seem less reliable. This can be explained in two ways, both based in the way both reliability measures are calculated. First, the estimated SSR_i was calculated using the SD of the so-called reference set. The rank-order SSR uses the SD of all representations (i.e. both the reference set and the new representation). This might lead to different results. A second explanation is that representations at the extremes of a rank order win or lose all their comparisons. This leads to a lack of information. It is unknown, respectively, how high or how low the actual ability score is. This results in a very inaccurate estimate, expressed in a high *se* value. This problem is commonly known as the separation problem in logistic regression (Kessels, Jones, and Goos 2013). The formula of the estimated SSR_i, as calculated in (**3**, **4**), only takes the *se* into account of one ability estimate, whereas the rank-order SSR, as calculated in (**6**, **7**), takes the average *se* into account over all ability estimates. Therefore, the estimated SSR_i for the extreme representations will be very small, thus lowering the average reported in Study 2.

GENERAL DISCUSSION

The current research addresses the efficiency problem of CJ assessments when a (semi-) random pair selection algorithm is used. Therefore, a newly developed adaptive pair selection algorithm was proposed and tested. This algorithm, inspired by computerized adaptive testing and based on a suggestion by Bramley (2015), Bramley and Vitello (2019), made use of a calibrated reference set of representations which functioned as a measuring scale for new representations. In a real-life assessment (Study 1) and a computer simulation (Study 2), it was examined if the adaptive algorithm was more efficient and more accurate in its parameter retrieval compared with a semi-random algorithm.

Overall, both studies show that in comparison with a semi-random selection algorithm, the reference set based adaptive selection algorithm produces more accurate results. The reference set-based adaptive selection algorithm is also more efficient than the semi-random selection algorithm, as it requires fewer comparisons (per representation) to reach a comparable level of reliability. Independent of the selection algorithm, the stopping criterion estimated SSR_i of 0.70 proves to be the most efficient (as measured by the number of comparisons per representation), whereas a fixed stopping criterion (10, 20, or 37 comparisons per representation) leads to the most accurate results (as measured by the RMSE). This shows that there is an efficiency–accuracy tradeoff. Based on the results of Study 2, it is advisable to use the estimated SSR_i of 0.90 as a stopping criterion. The consideration can be made that a difference in RMSE of 0.11 is worth an increase of seven comparisons per representation. This does, however, depend on the number of representations and the number of assessors available. Each practitioner needs to decide this for themselves. Finally, on the basis of Study 1, it can tentatively be concluded that it is recommended to use a reference set that is as accurate as possible.

It should be remarked, however, that in the adaptive algorithm the reliability is calculated for single estimates and not for a rank order. In Study 2, it was observed that for the semi-random selection algorithm on average, the estimated SSR_i is lower than the rank-order SSR. As already mentioned, these two reliability measures are, however, not completely comparable, namely, the formula for the estimated SSR_i uses the SD of the ability estimates in the reference set whereas the rank-order SSR uses the SD of all representations (also the new representations). These observations do not detract from the conclusion that the reference set based adaptive selection algorithm is more efficient and accurate than the semi-random selection algorithm.

There is also an effect of the accuracy of the reference set used in the adaptive algorithm. Study 1 shows that the higher the accuracy of the reference set, expressed by the rank-order SSR of the reference set, the higher the accuracy of the assessment results. However, from a reference set SSR of 0.80 on there appears to be no further gain in accuracy of the assessment results. Additionally, the difference in accuracy between the stopping criterion estimated SSR_i of 0.70 and the other two stopping criteria ($SSR_i = 0.90$ and $N_{CR} = 10$) decreases. This might be explained by the study setup. Because the order of the assessments was not counterbalanced between assessors, they might have been more tired when they reached the last session (with reference set SSR of 0.91). As a consequence, they might have made more mistakes. The adaptive algorithm does not allow to check if this is the case. In a semi-random algorithm, one could use the misfit statistics to see if an assessor made more errors in one assessment than in another[10] or if they made more mistakes than other assessors. Misfit statistics suppose that an assessor receives pairs of representations covering a broad range of ability differences[11] and that this range is approximately equal over assessors. This cannot be guaranteed with the adaptive algorithm described. In sum, it can be stated that with the current adaptive algorithm, a reference set with a reliability of at least 0.80 should be used.

Bramley and Vitello (2019) noted that adaptivity increases the spread of ability estimates. This means that adaptivity shifts ability estimates away from their true values. This is a problem when the ability estimates are used in a high stake assessment[12]. To check if the reference set based adaptive selection algorithm described in the current article suffers from the same issue, the standard deviation of the selected representations was calculated in the real assessment and for the generating values and the estimates in every replication of the simulation. The simulation study showed, contrary to Study 1, that there was no increase in standard deviation of the estimates in the reference set based adaptive selection algorithm compared with the generating values. This is probably because a reference set of representations with previously calibrated, and thus fixed, ability values are used. This helps to counter the missing information that adaptivity induces in other algorithms. However, it was observed that the standard deviation in the reference set based adaptive selection algorithm was higher than that in the semi-random selection algorithm when the more common stopping criteria (rank-order SSR and number of comparisons is 37) were used. This might be because in the simulation study a perfect assessor is used. Thus, collecting more comparisons might lead to a shift toward the mean for the results. Future research should look into this effect.

Some further remarks should be made regarding some limitations of this study and how these could be addressed by future research. First, it is recommended to use a reference set that is as accurate as possible (as expressed by its rank-order SSR).

[10] Taken that across the assessments the representations are of the same quality and the assessors are comparable.

[11] Difference in true ability of the two representations in a pair.

[12] A high stakes assessment is an assessment where the results are used to make important decisions for the person under assessment, e.g., a pass-fail decision.

However, because the order of the reference sets in Study 1 was not counterbalanced, it cannot be conclusively shown that the reference set reliability should either be 0.80, 0.90, or as high as possible. This should be confirmed in future research.

Second, the first study is only a single observation. Thus, it cannot be excluded that random errors influence the results. It might be informative to see what the range in accuracy and efficiency is when such an assessment is replicated over time (within assessor groups) and over assessor groups. This should further strengthen the tentative conclusions. By extension, it might be interesting to see how the results with the semi-random selection algorithm replicate over time and assessor groups. As far as we know, an extensive replication study of CJ assessment has not been conducted yet.

Third, the reference set reliability was not included in the simulation study in order to keep things feasible. Therefore, the results regarding the reference set reliability in Study 1 were not replicated here. A simulation study where reference sets of different reliability are constructed and used in the simulation with the reference set based adaptive selection algorithm might confirm the observations in Study 1 and check if the adaptive algorithm would benefit from an extended calibration study as is common in CAT.

Finally, some questions can be raised on what a sufficient number of representations in the reference set could be. Therefore, it is possible to look for inspiration in CAT because the reference set in the reference set based adaptive selection algorithm can be considered a resembling the test items in CAT. Therefore, in the number of representations in the reference set should be high enough in order to have a broad enough ability range and ability values close enough to each other to reach accurate results. On the other hand, the number of representations should be low enough that, when creating (or calibrating) the reference set, the work is still feasible for the assessors. The current research did not focus, however, on what this means in regard of specific numbers. Future research might thus look into how the number of representations in the reference set influence the performance of the reference set based adaptive selection algorithm. This can be done by comparing reference sets of different sizes and/or different ability ranges.

Disregarding these limitations, the adaptive algorithm as described in the current article shows it is more efficient compared with random pair construction. If users are willing to do an initial investment, the reference set could be used for multiple assessments in the future. It might even provide possibilities for standardized CJ assessments. This does, however, support on the assumption that CJ can be used to compare performances on different tasks as long as these tasks assess the same competency, because repeating the same assessment task year after year might encourage cheating and teaching to the test. To our knowledge, this assumption has not yet been investigated. Besides, exercises for standard maintaining across consecutive years in national assessment that are using CJ might also benefit from this adaptive algorithm. As already mentioned in the introduction, only one scale from a specific assessment year would be needed. This means a gain in time and effort in the next years. In both applications, however, techniques for updating and maintaining the reference set should be devised and tested.

DATA AVAILABILITY STATEMENT

The dataset analyzed in Study 1 cannot be made public, due to privacy restrictions in the "Contributors agreement" and related IB policy at the time of data collection. The data generated in Study 2 would produce very large files that are difficult to transfer. Because of IP restrictions, no working R code can be provided for the simulation. The results of every replication and the R code for the graphs and tables are freely available *via* the Zenodo repository: https://doi.org/10.5281/zenodo.5537019.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the International Baccaloreate, Legal and Compliance department. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

SV, AF, and RB contributed to the conception and design of Study 1. AF conducted Study 1, including recruiting and coaching assessors, selecting representations, and creating reference sets. SV provided technical support and set up all assessments from Study 1. SV designed and ran Study 2, developed the adaptive algorithm and conducted all analyses for Study 1 and 2. SV wrote the first draft of the manuscript. AF wrote sections of the manuscript. All authors contributed to manuscript revisions, read, and approved the submitted version.

FUNDING

This research was co-funded by a larger project (D-PAC) funded by the Flanders Innovation and Entrepreneurship and the Research Foundation (grant number 130043) and by the International Baccalaureate.

ACKNOWLEDGMENTS

The authors want to thank the assessors of Study 1 for their time an effort.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2021.785919/full#supplementary-material

REFERENCES

Bartholomew, S. R., Zhang, L., Bravo, E. G., and Strimel, G. J. (2019). A Tool for Formative Assessment and Learning in a Graphics Design Course: Adaptive Comparative Judgement. *Des. J.* 22 (1), 73–95. doi:10.1080/14606925.2018.1560876

Bouwer, R., Lesterhuis, M., Bonne, P., and De Maeyer, S. (2018). Applying Criteria to Examples or Learning by Comparison: Effects on Students' Evaluative Judgment and Performance in Writing. *Front. Educ.* 3, 86. doi:10.3389/feduc.2018.00086

Bradley, R. A. (1976). A Biometrics Invited Paper. Science, Statistics, and Paired Comparisons. *Biometrics* 32 (2), 213–239. doi:10.2307/2529494

Bradley, R. A., and Terry, M. E. (1952). Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 39 (3–4), 324–345. doi:10.1093/biomet/39.3-4.324

Bramley, T., Bell, J. F., and Pollitt, A. (1998). Assessing Changes in Standards over Time Using Thurstone Paired Comparisons. *Educ. Res. Perspect.* 25 (2), 1–24.

Bramley, T., and Gill, T. (2010). Evaluating the Rank-ordering Method for Standard Maintaining. *Res. Pap. Educ.* 25 (3), 293–317. doi:10.1080/02671522.2010.498147

Bramley, T. (2015). "Investigating the Reliability of Adaptive Comparative Judgment," in *Cambridge Assessment Research Report* (Cambridge, UK: Cambridge Assessment). Available at: www.cambridgeassessment.org.uk/Images/232694-investigating-the-reliability-of-adaptive-comparative-judgment.pdf.

Bramley, T. (2007). "Paired Comparison Methods," in *Techniques for Monitoring the Comparability of Examination Standards*. Editors P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, and P. Tymms (London, U.K: Qualifications and Curriculum Authority), 246–300.

Bramley, T., and Vitello, S. (2019). The Effect of Adaptivity on the Reliability Coefficient in Adaptive Comparative Judgement. *Assess. Educ. Principles, Pol. Pract.* 26, 43–58. doi:10.1080/0969594X.2017.1418734

Brignell, C., Wicks, T., Tomas, C., and Halls, J. (2019). The Impact of Peer Assessment on Mathematics Students' Understanding of Marking Criteria and Their Ability to Self-Regulate Learning. *MSOR Connections* 18 (1), 46–55. doi:10.21100/msor.v18i1.1019

Crompvoets, E. A. V., Béguin, A. A., and Sijtsma, K. (2020). Adaptive Pairwise Comparison for Educational Measurement. *J. Educ. Behav. Stat.* 45 (3), 316–338. doi:10.3102/1076998619890589

Curcin, M., Howard, E., Sully, K., and Black, B. (2019). "Improving Awarding: 2018/2019 Pilots," in *Research Repport Ofqual 19/6575. Research and Analysis* (Coventry, U.K: Ofqual). Available at: https://www.gov.uk/government/publications/improving-awarding-20182019-pilots.

Holmes, S. D., Meadows, M., Stockford, I., and He, Q. (2018). Investigating the Comparability of Examination Difficulty Using Comparative Judgement and Rasch Modelling. *Int. J. Test.* 18 (4), 366–391. doi:10.1080/15305058.2018.1486316

IBM Corp (2016). *IBM SPSS Statistics for Windows*. version 24.0. Armonk, NY: IBM Corp.

Jones, I., and Alcock, L. (2014). Peer Assessment without Assessment Criteria. *Stud. Higher Educ.* 39 (10), 1774–1787. doi:10.1080/03075079.2013.821974

Jonsson, A., and Svingby, G. (2007). The Use of Scoring Rubrics: Reliability, Validity and Educational Consequences. *Educ. Res. Rev.* 2 (2), 130–144. doi:10.1016/j.edurev.2007.05.002

Kessels, R., Jones, B., and Goos, P. (2013). "An Argument for Preferring Firth Bias-Adjusted Estimates in Aggregate and Individual-Level Discrete Choice Modeling," in *Research Report 2013* (Antwerp, Belgium: University of Antwerp, Faculty of Applied Economics). Working Papers.

Kimbell, R. (2021). Examining the Reliability of Adaptive Comparative Judgement (ACJ) as an Assessment Tool in Educational Settings. *Int. J. Techn. Des. Educ.* Online First. doi:10.1007/s10798-021-09654-w

Kimbell, R., Wheeler, T., Stables, K., Shepard, T., Martin, F., Davies, D., et al. (2009). *E-scape Portfolio Assessment Phase 3 Report*. London, U.K: Goldsmiths, University of London.

Laming, D. (2003). *Human Judgment: The Eye of the Beholder*. 1st ed. London, U.K: Cengage Learning EMEA.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York, NY: Wiley.

Molenaar, I. W. (1995). "3. Estimation of Item Parameters," in *Rasch Models: Foundations, Recent Developments, and Applications*. Editors G. H. Fischer and I. W. Molenaar (New York, NY: Springer-Verlag), 39–51.

Newhouse, C. P. (2014). Using Digital Representations of Practical Production Work for Summative Assessment. *Assess. Educ. Principles, Pol. Pract.* 21 (2), 205–220. doi:10.1080/0969594X.2013.868341

Nunnally, J. C. (1978). *Psychometric Theory*. 2nd ed. New York, NY: McGraw-Hill.

Pollitt, A., and Murray, N. L. (1995). "What Raters Really Pay Attention to," in *Studies in Language Testing 3: Performance Testing, Cognition and Assessment*. Editors M. Milanovic and N. Saville (Cambridge, U.K: Cambridge University Press), 74–91.

Pollitt, A. (2012). The Method of Adaptive Comparative Judgement. *Assess. Educ. Principles, Pol. Pract.* 19 (3), 281–300. doi:10.1080/0969594X.2012.665354

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. version 4.0.3. Vienna, Austria: R Foundation for Statistical Computing. Available at: www.R-project.org.

Thurstone, L. L. (1927). A Law of Comparative Judgment. *Psychol. Rev.* 34 (4), 273–286. doi:10.1037/h0070288

Thurstone, L. L. (1925). A Method of Scaling Psychological and Educational Tests. *J. Educ. Psychol.* 16 (7). doi:10.1037/h0073357

van Daal, T., Lesterhuis, M., Coertjens, L., van de Kamp, M.-T., Donche, V., and De Maeyer, S. (2017). The Complexity of Assessing Student Work Using Comparative Judgment: The Moderating Role of Decision Accuracy. *Front. Educ.* 2. doi:10.3389/feduc.2017.00044

Verhavert, S. (2018). "Chapter 2 Estimating the Bradley-Terry-Luce Model in R," in *Beyond a Mere Rank Order: The Method, the Reliability and the Efficiency of Comparative Judgment*. Available at: https://repository.uantwerpen.be/docman/irua/c24160/155690.pdf.

Verhavert, S., Bouwer, R., Donche, V., and De Maeyer, S. (2019). A Meta-Analysis on the Reliability of Comparative Judgement. *Assess. Educ.* 26 (5), 1–22. doi:10.1080/0969594x.2019.1602027

Verhavert, S., De Maeyer, S., Donche, V., and Coertjens, L. (2018). Scale Separation Reliability: What Does it Mean in the Context of Comparative Judgment. *Appl. Psychol. Meas.* 42 (6), 428–445. doi:10.1177/0146621617748321

Wright, B., and Stone, M. (1999). *Measurement Essentials*. 2nd ed. Wilmington, DE: Wide Range, Inc.

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed nor endorsed by the publisher.

frontiers in Education

ORIGINAL RESEARCH
published: 01 March 2022
doi: 10.3389/feduc.2021.788202

On the Bias and Stability of the Results of Comparative Judgment

Elise A. V. Crompvoets[1,2]*, *Anton A. Béguin*[3] *and Klaas Sijtsma*[1]

[1]Tilburg University, Tilburg, Netherlands, [2]Cito, Arnhem, Netherlands, [3]International Baccalaureate, Cardiff, United Kingdom

Comparative judgment is a method that allows measurement of a competence by comparison of items with other items. In educational measurement, where comparative judgment is becoming an increasingly popular assessment method, items are mostly students' responses to an assignment or an examination. For assessments using comparative judgment, the Scale Separation Reliability (SSR) is used to estimate the reliability of the measurement. Previous research has shown that the SSR may overestimate reliability when the pairs to be compared are selected with certain adaptive algorithms, when raters use different underlying models/truths, or when the true variance of the item parameters is below one. This research investigated bias and stability of the components of the SSR in relation to the number of comparisons per item to increase understanding of the SSR. We showed that many comparisons are required to obtain an accurate estimate of the item variance, but that the SSR can be useful even when the variance of the items is overestimated. Lastly, we recommend adjusting the general guideline for the required number of comparisons per item to 41 comparisons per item. This recommendation partly depends on the number of items and the true variance in our simulation study and needs further investigation.

Keywords: bias, comparative judgment (CJ), pairwise comparison (PC), reliability, stability

Edited by:
Renske Bouwer,
Utrecht University, Netherlands

Reviewed by:
Tom Benton,
Cambridge Assessment,
United Kingdom
Kaiwen Man,
University of Alabama, United States

***Correspondence:**
Elise A. V. Crompvoets
e.a.v.crompvoets@uvt.nl

Specialty section:
This article was submitted to
Assessment, Testing and Applied
Measurement,
a section of the journal
Frontiers in Education

Received: *01 October 2021*
Accepted: *17 December 2021*
Published: *01 March 2022*

Citation:
Crompvoets EAV, Béguin AA and Sijtsma K (2022) On the Bias and Stability of the Results of Comparative Judgment.
Front. Educ. 6:788202.
doi: 10.3389/feduc.2021.788202

INTRODUCTION

Comparative judgment is a method that allows measurement of a competence by comparison of items. When items are compared in pairs, comparative judgment is also known as pairwise comparison. This method has been used in different contexts ranging from sports to marketing to educational assessment, with different models for each context (e.g., Agresti, 1992; Böckenholt, 2001; Maydeu-Olivares, 2002; Maydeu-Olivares and Böckenholt, 2005; Böckenholt 2006; Stark and Chernyshenko, 2011; Cattelan, 2012; Brinkhuis, 2014). In educational measurement, where comparative judgment is becoming an increasingly popular assessment method (Lesterhuis et al., 2017; Bramley and Vitello, 2018), items are mostly students' responses to an assignment or an examination. The assignment or the examination is used to measure a competence of the students, and the students' responses give an indication of their competence level. The method has been used in a variety of contexts, ranging from art assignments (Newhouse, 2014) to academic writing (Van Daal et al., 2016) and mathematical problem solving (Jones & Alcock, 2013). These contexts have in common that the competencies are difficult to disentangle into sub-aspects together defining the competencies. Therefore, they are difficult to measure validly using analytical scoring schemes such as rubrics or criteria lists (Van Daal et al., 2016), which are conventional measurement methods used in education. In contrast to these analytic measurement methods, which assume that a competence can be operationalized by means of a list of sub-aspects and evaluate each aspect separately, comparative judgment is a holistic measurement method where a competence is evaluated as a whole (Pollitt, 2012); simply asking which of two items scores higher on the competence of interest suffices.

For complex competencies like art assignments, academic writing, and mathematical problem solving, it is possible that a higher validity can be obtained using comparative judgment instead of rubrics or criteria lists (Pollitt, 2012; Van Daal et al., 2016) because of its holistic character and the greater possibility of raters to use their expertise in their judgments compared to rubrics or criteria lists. In addition to the claim of higher validity of comparative judgment, Pollitt (2012) claimed that comparative judgment also results in higher reliability compared to using rubrics or criteria lists. However, later research has shown that this claim is likely to be too optimistic for the reported numbers of comparisons per item (e.g., Bramley, 2015; Bramley and Vitello, 2018; Crompvoets et al., 2020; Crompvoets et al., 2021), and that the extent to which high reliability that can be obtained using comparative judgment is limited (Verhavert et al., 2019).

To explain why Pollitt's (2012) claim is too optimistic, we first define two types of reliability in the context of comparative judgment: the benchmark reliability (Crompvoets et al., 2020, 2021) and the Scale Separation Reliability (SSR; e.g., Bramley, 2015; Crompvoets et al., 2020). Both forms of reliability are based on parameters of the Bradley-Terry-Luce (BTL; Bradley and Terry, 1952; Luce, 1959) model. This model is defined as follows. Let K be the number of items, let i and j ($i, j = 1, \ldots, K$) be item indices, and let θ_i and θ_j be the parameters of items i and j. Furthermore, let X_{ij} be the outcome of the inter-item comparison where $X_{ij} = 1$ means that item i was preferred to item j, and $X_{ij} = 0$ means that item j was preferred to item i. The BTL model defines the probability that item i is preferred to item j in a paired comparison by means of

$$P\left(X_{ij} = 1|\theta_i, \theta_j\right) = \frac{\exp\left(\theta_i - \theta_j\right)}{1 + \exp\left(\theta_i - \theta_j\right)}. \tag{1}$$

We interpret θ as an item parameter, but we may also interpret it as a person parameter for the competence of one person. For example, θ may represent the quality of a student's work, which in turn represents the competence level of the student. Thus, items and persons are not clearly distinguished in the BTL model for comparative judgment.

The benchmark reliability is only known in simulated data and is computed as the squared correlation between the true (simulated) item parameters and the item parameter estimates. Let θ be the item parameter in the generating model and let $\hat{\theta}$ be the item parameter estimate. The benchmark reliability can then be computed as

$$\rho_{\theta\hat{\theta}'} = \mathrm{cor}\left(\theta, \hat{\theta}\right)^2. \tag{2}$$

This definition of reliability corresponds with the definition of reliability as $\rho^2(\theta, \hat{\theta})$ in classical test theory (Lord and Novick, 1968), where θ represents the true score and $\hat{\theta}$ represents the observable test score. Since we are interested in reliability of the measurement of a specific set of items, benchmark reliability is used as the true reliability of this set of items.

The SSR is an estimate of reliability that is based on the Index of Subject Separation formulated by Andrich and Douglas (1977, as cited in Gustafsson, 1977) and is computed as follows. We assume that items are compared in pairs and that the location parameters of these items on the latent competence scale are of interest. Let $S^2(\theta)$ be the estimated true variance of the object parameters and let $S^2(\hat{\theta})$ be the variance of the estimated object parameters. Furthermore, let MSE be the mean of the squared standard errors corresponding to the item parameter estimates, computed as

$$MSE = \frac{1}{K}\sum_i^K SE\left(\hat{\theta}_i\right)^2.$$

The SSR can then be written as

$$SSR = \frac{S^2(\theta)}{S^2(\hat{\theta})} \tag{3}$$

where

$$S^2(\theta) = S^2(\hat{\theta}) - MSE,$$

that is, the observed variance minus an error term (Bramley, 2015).

Research (Bramley, 2015; Bramley and Vitello, 2018; Crompvoets et al., 2020) has shown that the SSR might overestimate reliability (**Eq. 2**) in certain situations. These include the use of certain adaptive algorithms to select the pairs that raters have to compare. Pollitt's (2012) claim that comparative judgment results in higher reliability than using rubrics or criteria lists is based on a study using an adaptive algorithm to select the pairs that are compared in combination with the SSR. Other situations in which the SSR may overestimate benchmark reliability are when raters behave inconsistent amongst each other, which would be reflected in the BTL model by different parameters for the same items, and perhaps when the true variance of the item parameters is below 1 as well (Crompvoets et al., 2021). The result that the SSR may overestimate reliability suggests why Pollitt's (2012) claim that comparative judgment results in higher reliability is likely too optimistic. Moreover, the result that the SSR may overestimate reliability is problematic because 1) reliability estimates should provide a lower bound to reliability to avoid reporting reliability that is too high and therefore promises too much (Sijtsma, 2009; Hunt and Bentler, 2015) and 2) most recommendations about the number of required comparisons are based on achieving at least a user-defined value of the SSR (e.g., Verhavert et al., 2019).

To the best of our knowledge, no one has thoroughly investigated and reported the positive bias of the SSR. Previous research that reported the bias of the SSR has stopped at the conclusion that the SSR was biased (Bramley, 2015; Bramley and Vitello, 2018) or has only led to speculations about the meaning of the bias due to either adaptive pair selection (Crompvoets et al., 2020), different rater probabilities, or small true variances (Crompvoets et al., 2021).

One might reason that the behavior of the SSR needs no investigation, because its value can easily be derived from the two components $S^2(\hat{\theta})$ and MSE (**Eq. 3**). The strategy to vary only one component and keep the other components constant shows how the value of the measure changes with the value of the component. However, both components of the SSR, $S^2(\hat{\theta})$ and MSE, are based on the parameter estimates $\hat{\theta}$ from the underlying model. This means that a shift in the item parameters affects both components simultaneously, which renders the strategy unrealistic for investigation of the SSR. In addition, all item parameter estimates are mutually dependent because we estimate the parameters based on comparisons of the items with each other. This means that every additional comparison changes all item parameter estimates, so we cannot vary one item parameter estimate keeping the other item parameter estimates constant. Moreover, the changes of item parameter estimates after one comparison depend on the parameters of the items that are compared; the outcome of the comparison, which is not always straightforward because we use a probabilistic model; the total number of items and their parameters; and the outcomes of all previous comparisons, which is not always straightforward due to the use of a probabilistic (e.g., BTL) model. In conclusion, instead of influencing the components of the SSR directly, we can only influence the set of item parameters, which influences the comparison data, which influences the parameter estimates, which influences the components of the SSR. Therefore, it is highly relevant to investigate the behavior of the SSR.

Because all quantities needed to estimate the SSR (**Eq. 3**) are based on the parameter estimates $\hat{\theta}$ from the underlying model, this study focused on the parameter estimates used in the computation of the SSR. Specifically, we investigated the bias and stability of the parameter estimates. We define these outcomes in the Method section. Because parameter estimates depend on the amount of data available, we investigated bias and stability of the parameter estimates in relation to the number of comparisons.

The goal of this study was to gain insight into the bias and stability of the parameter estimates and the SSR of comparative judgment in educational measurement from two perspectives. In addition, we aimed to use this information either to support the guideline about the number of required comparisons per item from Verhavert et al. (2019) or to provide a new guideline based on the results from this study. First, we adapted the guideline for the required number of observations to obtain stable results for the one-parameter item response model or Rasch model (Rasch, 1960) for regular multiple choice tests to the BTL model used for comparative judgment. Second, we investigated the bias and stability of the parameter estimates and SSR of comparative judgment in a simulation study. In the discussion, we will reflect on the two perspectives.

SAMPLE SIZE GUIDELINE ADAPTATION TO THE BRADLEY-TERRY-LUCE MODEL

To determine the required number of observations to obtain stable model parameters, most researchers and test institutions use experience as their guide. One reason for this may be that the literature about sample size requirements to obtain stable model parameters is sparse and seems limited to conference presentations (Parshall et al., 1998), articles that were not subjected to peer review (Linacre, 1994), a framework used to assess test quality written in a non-universal language (Evers et al., 2009), or a brief mention in a book (Wright and Stone, 1979, *p*. 136). Parshall et al. (1998) and Evers et al. (2009) describe the guideline that for the one-parameter item response model, at least 200 observations per item are required to obtain stable item location parameter estimates. Wright and Stone (1979) suggest using 200 observations for test linking using the Rasch model, although they, and Linacre (1994), also mention that fewer observations may be sufficient to obtain sufficiently stable parameter estimates for some purposes. When the model parameters are considered sufficiently stable depends on the context. Because we encountered the guideline of 200 observations per item for several purposes and it is used often in practice, we used this guideline as a starting point.

The literature about guidelines for the Rasch model may be sparse, but for the mathematically related (Andrich, 1978) BTL model, no guidelines exist that describe how many observations are required in educational measurement for obtaining stable item parameter estimates. In this section, we first describe how the Rasch model and the BTL model are related, and then adapt the guideline from the Rasch model to the BTL model. In the Discussion section, we will evaluate this guideline in relation to the outcomes of the simulation study from the next section and in relation to the literature.

The Rasch model is defined as follows. Let N be the number of persons in the sample, let i ($i = 1, \ldots, N$) be the person index, and let θ_i^* be the parameter of person i on the latent variable scale, where the $*$ indicates that θ_i^* differs from θ_i used in the BTL model (**Eq. 1**). Let K be the number of items, let j ($j = 1, \ldots, K$) be the item index, and let β_j be the parameter of item j on the latent variable scale. Furthermore, let X_{ij} be the outcome of the person-item comparison where $X_{ij} = 1$ means that person i answered item j correctly, and $X_{ij} = 0$ means that person i answered item j incorrectly. The Rasch model defines the probability that person i answers item j correctly by means of

$$P\left(X_{ij} = 1|\theta_i^*, \beta_j\right) = \frac{\exp\left(\theta_i^* - \beta_j\right)}{1 + \exp\left(\theta_i^* - \beta_j\right)}. \tag{4}$$

We note that although mathematically it would have made sense to use β_i and β_j in the formulation of the BTL model (**Eq. 1**) for equivalence with the Rasch model, we chose to follow the conventional notation of the BTL model in comparative judgment contexts using θ notation for the items.

Even though the Rasch model and the BTL model have different parametrization (Verhavert et al., 2018), Andrich (1978) showed that the equations for the Rasch model and the BTL model are equivalent. This means that a person-item comparison in the Rasch model is mathematically equivalent to an inter-item comparison in the BTL model. Therefore, it makes sense to adapt the guideline for the Rasch model about the required number of observations for stable model estimates to the BTL model.

Our starting point for the guideline adaptation is the item, since items are present in both the Rasch model and the BTL model. In addition, the guideline Parshall et al. (1998) suggested aims at obtaining stable item parameter estimates. We assume that the number of items in the test that the Rasch model analyzes is the same as the number of items in the set of paired comparisons that is analyzed by means of the BTL model. However, the manner in which we obtain additional observations for an item differs between the models. Each observation for an item in the Rasch model is obtained from a person belonging to a population with many possible parameter values, whereas each observation for an item in the BTL model is obtained from an item in the fixed set of items under investigation. Therefore, for the BTL model, the information obtained from one observation may depend on the item parameters in the set, which is different for the Rasch model, where the information also depends on the sample of persons.

There are two ways to adapt the guideline from the Rasch model for use with the BTL model. The first adaptation is to equate the number of required observations per item for the BTL to the required number for the Rasch model; that is, 200 observations per item (Guideline 1). Since each comparative judgment/observation for the BTL model contains information about two items, this adaptation means that compared to the Rasch model, we need half of the total number of observations. We illustrate this with an example. Suppose we have 20 items in both models. The guideline of 200 observations per item for the Rasch model means that we need 200 (persons) × 20 (items) = $4,000$ observations in total for a 20-item test to obtain stable item parameter estimates. The adapted guideline of 200 observations per item for the BTL model means that we need $200\ (\text{comparisons per item}) / 2\ (\text{items per comparison}) \times 20\,(\text{items}) = 2000$ observations in total for a 20-item test to obtain stable item parameter estimates.

The second possibility is to equate the total number of observations for a set of items instead of the number of observations of one item. Continuing the example from the previous paragraph, 4,000 observations are required for a set of 20 items for the Rasch model to obtain stable item parameter estimates using Parshall et al.'s (1998) guideline. Adapted to the BTL model following the second guideline (i.e., equating the total number of observations for a set of items), this would mean that 4,000 paired comparisons in total are required to get stable item parameter estimates, which would mean $(4,000\ \text{comparisons} \times 2\ \text{items per comparison}) / 20\ \text{items}\ = 400$ observations per item. This is our Guideline 2. This means that compared to the Rasch model, we need twice as many observations per item for stable item parameter estimates from the BTL model. This makes sense, because each observation in a comparative judgment setting contains information about two items, so only half of the information concerns each item. We will evaluate both guidelines in the discussion section of this paper. One should note that the current recommendations for the numbers of comparisons per item based on a meta-analysis of comparative judgment applications range from 12 to 37 (Verhavert et al., 2019), which shows a large discrepancy with both the 200 and 400 comparisons per item according to the two adapted guidelines.

For the BTL model, the limited number of unique comparisons implies that the number of items in the set influences which numbers are compared, even though the number of observations per item does not change for different numbers of items. The number of items in the set is nonlinearly related to the number of unique comparisons in a comparative judgment setting. This means that the number of times each unique comparison is made differs for different numbers of comparisons. **Table 1** illustrates this: using guideline 2, for 20 items, all unique comparisons should be made 21 times (on average). On the other hand, for 1,000 items, all unique comparisons should be made 0.4 times, which means that not even all unique comparisons are made.

BIAS AND STABILITY OF SCALE SEPARATION RELIABILITY COMPONENTS

We investigated in a simulation study: 1) How many comparisons are required to obtain a stable and unbiased variance of the parameter estimates, $S^2(\hat{\theta})$; 2) how many comparisons are required to obtain a coverage of 95%-confidence intervals for the parameter estimates $\hat{\theta}$ using the standard errors $\mathrm{SE}(\hat{\theta})$ of 95%; and 3) how the SSR develops with increasing number of comparisons. We investigated these outcomes in situations in which we expected the SSR to underestimate benchmark reliability, because it is easier to understand the SSR and its components in these situations than in situations where we do not know why the SSR overestimates benchmark reliability. The R-code of the simulation study is available at https://osf.io/x7qzc/.

METHODS

Simulation Set-Up

The simulation design had two factors. First, we varied the number of items $N = \{20, 30, 50, 100\}$ to investigate whether the number of items affects the stability of the SSR estimate. Second, we used five different variances of the simulated item parameters. In the first condition, we used a variance of zero, which means that all items had the same location on the scale. We used this condition as a benchmark to investigate when the SSR was stable at zero, because the SSR should be zero if the true variance is zero, see (**Eq. 3**). In the second condition, we used a variance of 1.59, which is a realistic value based on the

TABLE 1 | Total number of observations and number of complete designs according to the translated guideline for the BTL model for different numbers of items.

Number of items	20	50	100	200	500	1,000
			Guideline 1: 200 observations per item			
Total number of observations	2,000	5,000	10,000	20,000	50,000	100,000
Number of complete designs[a]	10.53	4.08	2.02	1.01	0.40	0.20
			Guideline 2: 400 observations per item			
Total number of observations	4,000	10,000	20,000	40,000	100,000	200,000
Number of complete designs[a]	21.05	8.16	4.04	2.01	0.80	0.40

[a] *One complete design contains all $N(N-1)/2$ unique comparisons.*

argumentative writing dataset ‘having children’ used in Van Daal (2020, data retrieved from https://osf.io/wpbhk/?view_only=7aa609162ca146bbbbe9236c9224b668). Argumentative writing refers to one’s ability to express, argue for, and refute objections of one’s opinion about a specific topic (Van Daal, 2020, *p*. 175). This dataset contained 1,224 comparative judgments performed by 55 raters of 135 texts written by students in the fifth year of secondary education on the topic ‘having children’. Based on a comparison with the summary of several datasets in the meta-analysis of Verhavert et al. (2019), we argue that this dataset is realistic and representative of datasets obtained using comparative judgment for educational measurement. Furthermore, we added the variance conditions 0.5, 1, and 3 to obtain information about the results in between and beyond the benchmark variance and the realistic variance.

For each of the 4 (Number of Items) x 5 (Variance of Items) = 20 design conditions, we repeated the same procedure 100 times. We first selected to item pairs (1,2) (2,3), (3,4), et cetera, until $(K-1, K)$ and $(K,1)$ to create a linked comparison design. For each item pair, we simulated a comparison in which the probability of preferring one item to the other was given by the BTL model (**Eq. 1**). After these K comparisons, we estimated the BTL model using the open-source R-code from Crompvoets et al., 2020. This code uses an Expectation Maximization algorithm based on Hunter (2004) to obtain Maximum Likelihood estimates of the parameters. We used the parameter estimates from the BTL model to compute $S^2(\hat{\theta})$ and the SSR for the first time. Subsequently, we compared a randomly selected pair of items, estimated the BTL model parameters, and computed $S^2(\hat{\theta})$ and the SSR after each comparison until the maximum number of comparisons of 200 per item was reached. Lastly, we computed the number of comparisons per item required to obtain a stable variance of the parameter estimates $S^2(\hat{\theta})$ at the true parameter variance and the number of comparisons per item required to obtain a correct coverage of the 95% confidence interval for the parameter estimates $\hat{\theta}$.

We determined the number of comparisons per item required for a stable and accurate estimate to be the number of comparisons where $12K$ subsequent comparisons produced a value within a range around the true value, both for $S^2(\hat{\theta})$ and for the coverage of the 95% confidence intervals. The range of accurate values was defined as the range between 1 standard error below the true value and 1 standard error above the true value. We based the $12K$ subsequent comparisons on the guideline of 12 comparisons per item from the meta-analysis of Verhavert et al. (2018).

RESULTS

Figure 1 shows the development of $S^2(\hat{\theta})$ (top row), MSE (middle row), and the SSR (bottom row) with increasing numbers of comparisons per item of each of the 100 simulations per design cell and of the average for all true variance conditions for 50 items. On average, $S^2(\hat{\theta})$ seems to converge to the true variance, but not for every single simulated data set. Comparing the top- and middle rows, we see that there is much more variation in $S^2(\hat{\theta})$ than in MSE across simulations. The variation in development across simulations of both $S^2(\hat{\theta})$ and MSE was larger for larger true variance values. Interestingly, although $S^2(\hat{\theta})$ and MSE are the only components needed to compute the SSR (**Eq. 2**), the variation in development across simulations of the SSR shows the opposite trend with smaller variation for larger true variance values.

Figure 2 shows the development of bias in $S^2(\hat{\theta})$ (top row) and bias in SSR (bottom row) with increasing numbers of comparisons per item averaged across all 100 simulations with 68% confidence intervals for both true variance conditions and all numbers of items. In general, the bias of $S^2(\hat{\theta})$ was smaller for larger numbers of items. We first describe the results for a true variance of 0. The bias of $S^2(\hat{\theta})$ was larger for smaller numbers of items, but differences in $S^2(\hat{\theta})$ among numbers of items almost disappeared after about 30 comparisons per item. For 20 and 30 items, the SSR overestimated benchmark reliability in the beginning of data collection. For 20 items, this overestimation stopped after only a few comparisons, but then underestimated benchmark reliability by about 0.2 units. For 30 items, it took about 25 comparisons per item to stop the SSR from overestimating benchmark reliability. For 50 items, the SSR closely estimated benchmark reliability after only a few comparisons per object. For 100 items, the SSR closely estimated benchmark reliability after about 40 comparisons.

We next describe the results for the other true variances. In general, the differences among the number of items conditions in $S^2(\hat{\theta})$ were larger for larger true variances. For true variances larger than 1, on average, $S^2(\hat{\theta})$ was underestimated for 100 items, while it was overestimated for lower numbers of items and lower true variances. Except for a true variance of 3, fewer comparisons were required to converge to the true variance for larger

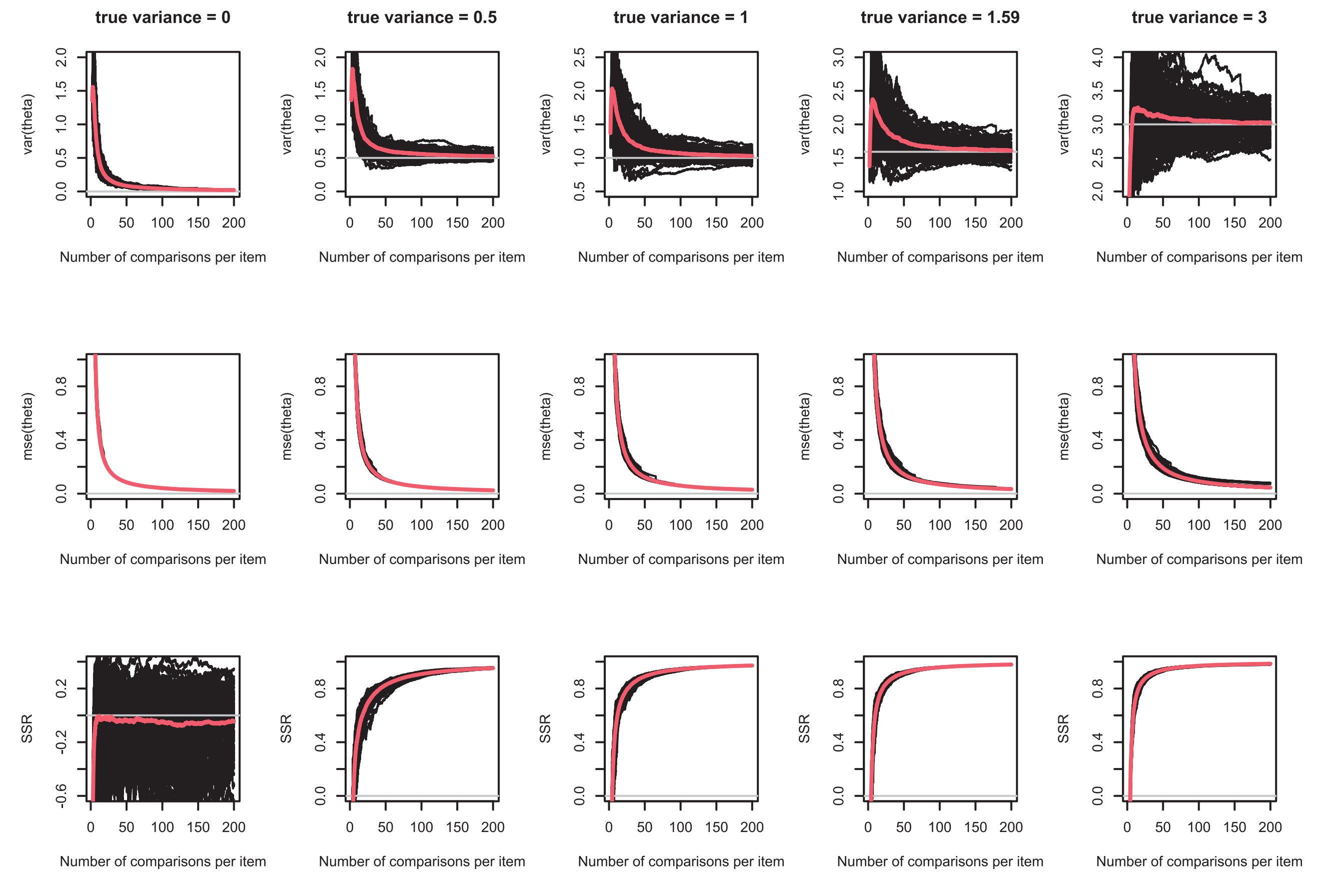


FIGURE 1 | Development of $S^2(\hat{\theta})$, $MSE[S(\hat{\theta})]$, and the SSR with increasing numbers of comparisons per item of all 100 simulations (black lines) and of the average (red line) for all true variance conditions for 50 items. The *x*-axis shows the average number of comparisons per item for interpretation purposes, but the data points are per comparison.

numbers of items. The SSR closely estimated benchmark reliability often after a few comparisons but almost always with 30 comparisons per item. Furthermore, on average, the SSR seemed to closely estimate benchmark reliability after fewer comparisons for lower numbers of items, which is the opposite trend of convergence compared to $S^2(\hat{\theta})$. However, the differences in SSR among the numbers of items are quite small in general. One difference worth mentioning is that for 20 items and a true variance of 0.5, the SSR was overestimated in the beginning of data collection, which is more like the condition with a true variance of zero.

Table 2 shows the mean number of comparisons per item required for accurate $S^2(\hat{\theta})$ values. In general, fewer comparisons per item are required on average for larger numbers of items, with the exception of 100 items and a true variance of 3. In addition, more comparisons per item are required on average for increasing true variances, with the exception of 100 items and a true variance of 3. The mean number of comparisons per item required for accurate $S^2(\hat{\theta})$ values ranges from 24 comparisons per item (for 100 items and a true variance of 1.59) to 119 comparisons per item (for 20 items and a true variance of 0.5). Furthermore, the large ranges within each condition indicate that there is a large variation in the number of comparisons per item required across simulations.

Figure 3 shows the development of the coverage of the 95% confidence intervals for the parameter estimates $\hat{\theta}$ with increasing numbers of comparisons per item. In general, with the exception of 100 items and a true variance of 3, the coverage was larger than 95%, which indicates that the standard errors of the parameter estimates were overestimated. However, most values are within the range of accurate values. The number of items required for accurate coverage was lower for larger true variances (**Figure 3**; **Table 3**). As **Table 3** indicates, in many conditions, the coverage was accurate in 12 comparisons per item or under, and it was accurate for at most 25 comparisons per item.

Because the development of $S^2(\hat{\theta})$ and the coverage with increasing number of comparisons per item was different from the development of the SSR, we decided to provide a guideline based on the SSR itself instead of its components. To this end, we computed the number of comparisons per item required for the SSR to underestimate benchmark reliability within a margin in 95% of the cases. Specifically, we calculated how many

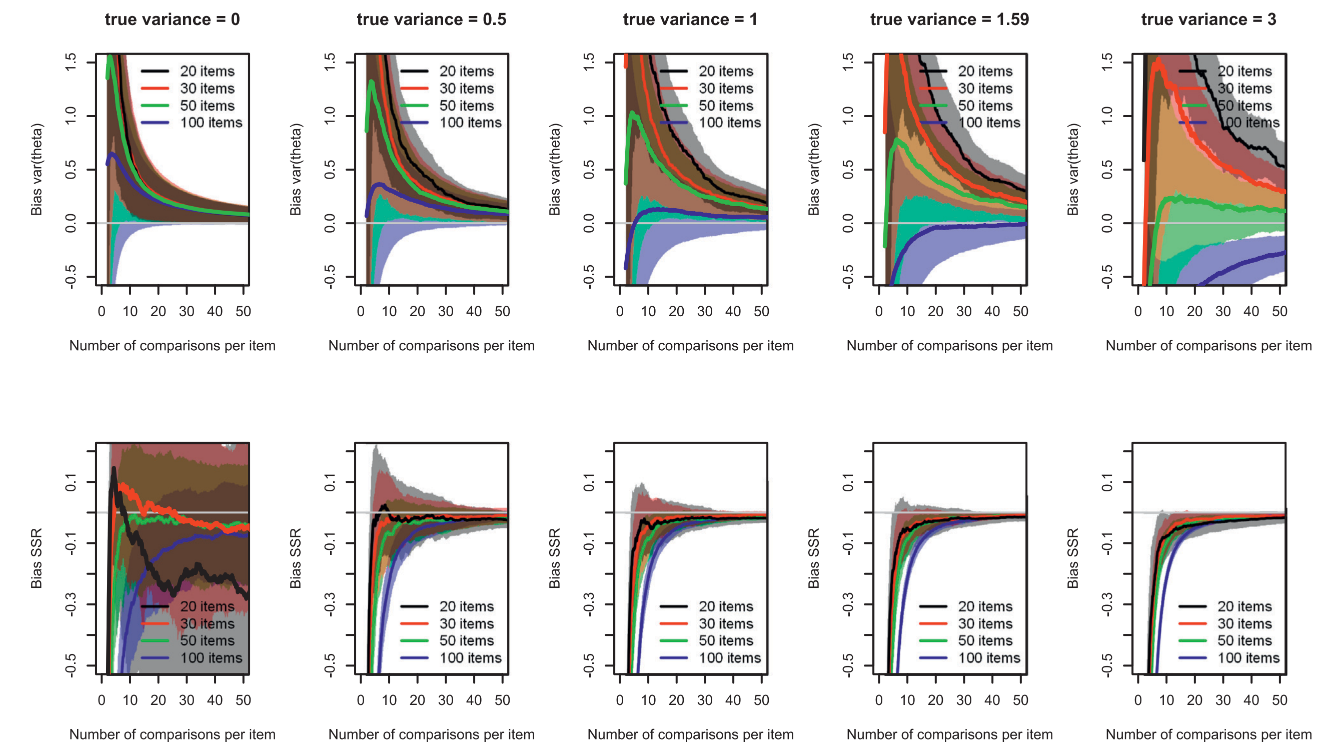


FIGURE 2 | Development of bias in $S^2(\hat{\theta})$ and bias in SSR with increasing numbers of comparisons per item averaged across simulations with 68% confidence intervals for all true variance conditions and all numbers of items. Bias in SSR was computed as *SSR—benchmark reliability*. For $S^2(\hat{\theta})$, we used $MSE[S(\hat{\theta})]$ to create the confidence interval. For the SSR, we used the SD across simulations to create the confidence interval. The *x*-axis shows the average number of comparisons per item (up to 50 comparisons per item) for interpretation purposes, but the data points are per comparison.

TABLE 2 | Mean number of comparisons per item required for accurate estimation of the true variance.

Number of items	True variance *M (min-max)*[a]			
	0.5	1.0	1.59	3.0
20	119 (29–200+)	100 (18–200+)	102 (21–200+)	88 (14–200+)
30	98 (18–200+)	78 (13–200+)	69 (13–200+)	68 (13–200+)
50	93 (13–200+)	68 (14–200+)	50 (10–200+)	42 (5–161)
100	72 (18–200+)	31 (4–121)	24 (6–94)	54 (14–200+)

[a] *Based on 100 simulations.*

Note. The number of comparisons per item represents the average number of comparisons per item in a set of items (i.e., one item may be compared more often than another item) rounded up to integers.

comparisons per item were required such that the lower bound of the 95% CI of the SSR was between the benchmark reliability and a margin of 0.10, 0.05, 0.03, and 0.01 below the benchmark reliability for each condition. The results are displayed in **Table 4**. The number of comparisons per item required for the SSR to closely estimate benchmark reliability depended on the number of items in the set and the true variance of the item parameters, which is in line with the bottom row in **Figures 1**, **2** displaying the SSR in relation to the number of comparisons per item. The number of comparisons per item ranged from 15 to more than 200. In general, smaller margins led to more comparisons per item required, more items in a set led to approximately the same or fewer comparisons per item required, and larger true variances led to fewer comparisons per item required, except for the combination of 20 items and a true variance of 3.

DISCUSSION

The guideline that 200 observations per item are required for stable parameter estimates using the Rasch model (Parshall et al., 1998) was adapted for the BTL model in two ways. Guideline 1 was obtained using the number of observations per item in the Rasch model, resulting in 200 comparisons per item for the BTL model. Guideline 2 was obtained using the total number of

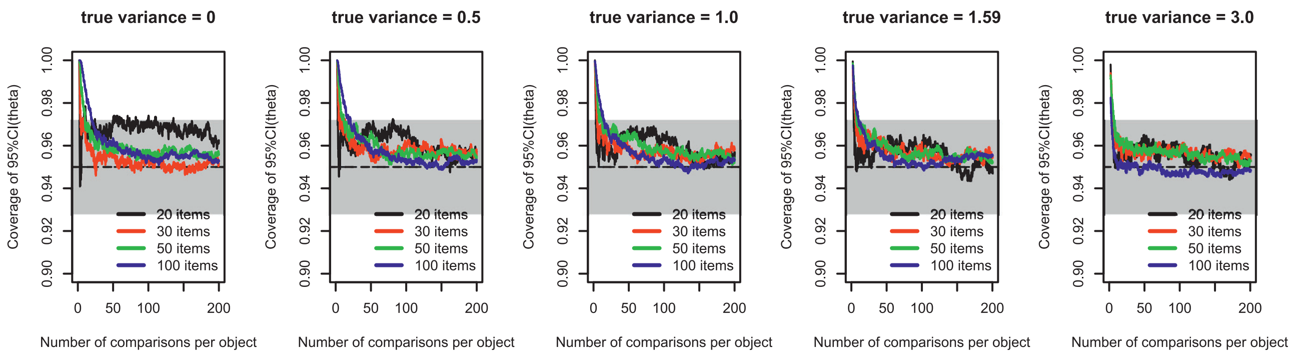


FIGURE 3 | Development of the coverage of 95% confidence intervals for parameter estimates $\hat{\boldsymbol{\theta}}$ with increasing numbers of comparisons per item for all true variance conditions and all numbers of items.

TABLE 3 | Mean number of comparisons per item required for accurate coverage of 95% CI around parameter estimates.

Number of items	True variance				
	0	0.5	1.0	1.59	3.0
20	25	4	4	4	5
30	9	14	6	6	7
50	11	12	12	11	7
100	21	21	16	11	4

Note. The number of comparisons per item represents the average number of comparisons per item in a set of items (i.e., one item may be compared more often than another item) rounded up to integers.

observations in a set of items in the Rasch model, resulting in 400 comparisons per item for the BTL model.

In the simulation study, the results showed that the variation in development across simulations of both the estimated variance and the mean squared standard error were larger for larger true variance values, but the variation in development across simulations of the SSR was smaller for larger true variance values. This is interesting, because the estimated variance and the mean squared standard error are the only components of the SSR. Possibly, the variations in the estimated variance and the mean squared standard error are more aligned for larger true variances such that combining them in the SSR leads to less variation. On average, the variance was accurately estimated after 24 to 119 comparisons per item, although the number of comparisons per item differed greatly among simulations. The coverage of the 95% confidence intervals of the parameter estimates showed that the standard errors of the parameter estimates were accurate after 4 to 25 comparisons per item. The SSR could closely estimate benchmark reliability even when the variance of the parameter estimates was still overestimated. When using margins ranging from 0.10 to 0.01 to determine when the SSR closely estimated benchmark reliability, across conditions, the number of comparisons per item ranged from 15 to more than 200.

When we compare the results from the two perspectives, it seems that Guideline 2 of 400 comparisons per item is too pessimistic and overly demanding. Guideline 1 could be useful since several simulations took 200 or more comparisons per item to get stable variance estimates and it took 200 or more comparisons for the SSR to closely estimate benchmark reliability when the margin was 0.01. However, averaged across samples, the variance was accurately estimated after a maximum of 119 comparisons per item, the standard errors of the parameters and the SSR required even fewer comparisons per item, and in most conditions, the SSR closely estimated benchmark reliability after less than 50 comparisons per item. Therefore, Guideline 2 may be too demanding as well.

TABLE 4 | Number of comparisons per item required for the SSR to estimate benchmark reliability between the benchmark reliability value and the benchmark reliability value minus the margin in 95% of the cases.

Number of items	Margin			
	0.10	0.05	0.03	0.01
	True variance = 0.5			
20	**41**	72	112	200+
30	**32**	62	97	200+
50	**32**	57	75	175
100	**28**	45	58	105
	True variance = 1			
20	27	48	70	136
30	19	33	49	135
50	18	36	53	108
100	19	30	42	77
	True variance = 1.59			
20	23	42	59	119
30	17	29	42	97
50	16	25	39	78
100	18	28	37	69
	True variance = 3			
20	25	58	100	200+
30	16	27	43	113
50	15	22	36	83
100	17	25	33	64

Note. The number of comparisons per item represents the average number of comparisons per item in a set of items (i.e., one item may be compared more often than another item) rounded up to integers. Underline for advised (maximum) number of comparisons per item for each threshold. Bold for advised (maximum) number of comparisons per item for each number of items.

The alternative guideline we present here is largely based on **Table 4**. We recommend that comparative judgment applications require at least 41 comparisons per item based on the following considerations. In general, smaller margins led to more comparisons per item required, more items in a set led to approximately the same or fewer comparisons per item required, and larger true variances led to fewer comparisons per item required. With respect to the margin that determines how much the SSR may underestimate benchmark reliability, we are lenient by choosing the largest margin. We believe that this is justified because the benchmark reliability is usually larger than the SSR, and because Verhavert et al. (2019) indicate that the SSR already has high values with this many comparisons per item. If one prefers a smaller margin, we recommend 72 comparisons per item for a margin of 0.05, 112 comparisons for a margin of 0.03, and more than 200 comparisons for a margin of 0.01. With respect to the true variance of the item parameters, we were quite strict by choosing the largest number of comparisons, which was for a true variance of 0.5. Because one can never know the true variance in practice and because our study showed that accurate variance estimation often required many observations per item, we argue that it is best to play safe, that is, to risk performing more comparisons than required for the desired accuracy rather than risking that you do not achieve the desired accuracy by performing too few comparisons. For example, if the number of comparisons for a comparative judgment application is based on a variance of 1, but in reality the true variance is less than 1, the SSR will not be as close to the benchmark reliability as one may believe. With respect to the number of items, we also argue to be strict and play safe. Therefore, we chose the number of comparisons for 20 items for the general guideline, which requires the most comparisons per item. However, as one does know the number of items in their comparative judgment application, the required number of comparisons can be somewhat adjusted to the number of items in this set. **Table 4** provides information about this adjustment, but the researcher must make the call, given that we only investigated four numbers of items.

Our guideline of 41 comparisons per item renders comparative judgment less interesting to use in practice than the guideline of 12 comparisons per item Verhavert et al. (2019) suggested. However, 41 comparisons per item are necessary for accurately determining the reliability of the measurement using the SSR. The SSR may overestimate benchmark reliability in individual samples, even when it underestimates reliability on average, especially when the number of comparisons is small. Based on **Table 4**, we suggest that after 41 comparisons, the risk of overestimating reliability with the SSR in individual samples is largely reduced.

Our guideline concerns reliability estimation by means of the SSR and not benchmark reliability. This means that using fewer than 41 comparisons may result in sufficient benchmark reliability (Crompvoets et al., 2020; Crompvoets et al., 2021). The problem is that we cannot determine whether this is the case based on the SSR. Therefore, if a different reliability estimate would exist for comparative judgment, the guideline might change. Measures like the root mean squared error (RMSE) may be useful in some instances, since it is related to reliability, only in terms of the original scale. However, the fact that the RMSE is scale dependent also makes it more difficult to interpret and to compare between different measurements. Therefore, a standardized measure of reliability, bound between 0 and 1, would be preferred. This is an interesting topic for future research.

In our simulation designs, we did not use adaptive pair selection algorithms or multiple raters who perceived a different truth, which are the situations in previous research where the SSR systematically overestimated benchmark reliability. The results of our study provide a baseline how the SSR and the components used to compute the SSR develop with increasing numbers of comparisons when the SSR is expected to underestimate reliability, as it should. Future research could build on our results by investigating how the components of the SSR develop with increasing numbers of comparisons in situations where the SSR might overestimate reliability. The fact that the SSR might overestimate reliability in some situations is even more reason to use a guideline that reduces the risk of overestimation due to sampling fluctuations.

Our study focused on the components of the SSR because we expected that this would show the cause of the inflation of the SSR. However, our simulation study showed that the estimated variance and standard errors of the item parameters developed differently from the SSR with increasing numbers of comparisons with respect to variation between samples, which is not what we expected. Since the components of the SSR developed differently from the SSR, they do not seem to be the cause of the inflation of the SSR. Future research could also aim at developing alternative reliability estimates to the SSR.

In conclusion, the SSR may overestimate reliability in certain situations, but it can function correctly as an underestimate of reliability even when the variance of the items is overestimated. The SSR can be used when the pairs to be compared are selected without an adaptive algorithm, when raters use the same underlying model/truth, and when the true item variance is at least 1. The variance of the items is likely to be overestimated when fewer than 24 comparisons per item were performed. An adaptation of the guideline for the Rasch model was too pessimistic. We provided a new guideline of 41 comparisons per item, with nuances concerning the number of items and the margin of accuracy for SSR estimation. Future research is needed to further investigate the SSR estimation and to develop an alternative reliability estimate.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://osf.io/x7qzc/.

AUTHOR CONTRIBUTIONS

EC executed the research and wrote the manuscript. AB and KS contributed to the analysis plan and writing of the manuscript.

REFERENCES

Agresti, A. (1992). Analysis of Ordinal Paired Comparison Data. *Appl. Stat.* 41, 287–297. doi:10.2307/2347562

Andrich, D. (1978). Relationships between the Thurstone and Rasch Approaches to Item Scaling. *Appl. Psychol. Meas.* 2, 451–462. doi:10.1177/014662167800200319

Böckenholt, U. (2001). Thresholds and Intransitivities in Pairwise Judgments: A Multilevel Analysis. *J. Educ. Behav. Stat.* 26, 269–282. doi:10.3102/10769986026003269

Böckenholt, U. (2006). Thurstonian-based Analyses: Past, Present, and Future Utilities. *Psychometrika* 71, 615–629. doi:10.1007/s11336-006-1598-5

Bradley, R. A., and Terry, M. E. (1952). Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 39, 324–345.

Bramley, T. (2015). *Investigating the Reliability of Adaptive Comparative Judgement*. Cambridge Assessment Research Report. Cambridge, United Kingdom: Cambridge Assessment.

Bramley, T., and Vitello, S. The Effect of Adaptivity on the Reliability Coefficient in Adaptive Comparative Judgement. *Assess. Educ. Principles, Pol. Pract.* (2018), 26, 43–58. doi:10.1080/0969594X.2017.1418734

Brinkhuis, M. J. S. (2014). *Tracking Educational Progress*. Amsterdam, Netherlands: Doctoral dissertation, University of Amsterdam. Retrieved from: https://pure.uva.nl/ws/files/2133789/153696_01_1_.pdf.

Cattelan, M. (2012). Models for Paired Comparison Data: A Review with Emphasis on Dependent Data. *Statist. Sci.* 27, 412–433. doi:10.1214/12-STS396

Crompvoets, E. A. V., Béguin, A. A., and Sijtsma, K. (2020). Adaptive Pairwise Comparison for Educational Measurement. *J. Educ. Behav. Stat.* 45, 316–338. doi:10.3102/1076998619890589

Crompvoets, E. A. V., Béguin, A. A., and Sijtsma, K. (2021). *Pairwise Comparison Using a Bayesian Selection Algorithm: Efficient Holistic Measurement*. Available at: https://psyarxiv.com/32nhp/.

Evers, A., Lucassen, W., Meijer, R. R., and Sijtsma, K. (2009). COTAN beoordelingssysteem voor de kwaliteit van tests *[COTAN assessment system for the quality of tests]*. Amsterdam, The Netherlands: Nederlands Instituut van Psychologen.

Gustafsson, J.-E. (1977). *The Rasch Model for Dichotomous Items: Theory, Applications and a Computer Program*. Göteborg, Sweden: Göteborg University.

Hunt, T. D., and Bentler, P. M. (2015). Quantile Lower Bounds to Reliability Based on Locally Optimal Splits. *Psychometrika* 80, 182–195. doi:10.1007/s11336-013-9393-6

Hunter, D. R. (2004). MM Algorithms for Generalized Bradley-Terry Models. *Ann. Statist.* 32, 384–406. doi:10.1214/aos/1079120141

Jones, I., and Alcock, L. (2013). Peer Assessment without Assessment Criteria. *Stud. Higher Edu.* 39, 1774–1787. doi:10.1080/03075079.2013.821974

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., and De Maeyer, S. (2017). "Comparative Judgement as a Promising Alternative to Score Competences," in *Innovative Practices for Higher Education Assessment and Measurement* (Hershey, PA: IGI Global), 119–138. doi:10.4018/978-1-5225-0531-0.ch007

Linacre, J. M. (1994). Sample Size and Item Calibration Stability. *Rasch Meas. Trans.* 7, 328, 1994 . Retrieved from: https://rasch.org/rmt/rmt74m.htm.

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Boston, Massachusetts, United States: Addison-Wesley.

Luce, R. D. (1959). *Individual Choice Behaviours: A Theoretical Analysis*. New York, NY: Wiley.

Maydeu-Olivares, A., and Böckenholt, U. (2005). Structural Equation Modeling of Paired-Comparison and Ranking Data. *Psychol. Methods* 10, 285–304. doi:10.1037/1082-989X.10.3.285

Maydeu-Olivares, A. (2002). Limited Information Estimation and Testing of Thurstonian Models for Preference Data. *Math. Soc. Sci.* 43, 467–483. doi:10.1016/s0165-4896(02)00017-3

Newhouse, C. P. (2014). Using Digital Representations of Practical Production Work for Summative Assessment. *Assess. Educ. Principles, Pol. Pract.* 21, 205–220. doi:10.1080/0969594X.2013.868341

Parshall, C. G., Davey, T., Spray, J. A., and Kalohn, J. C. (1998). Computerized Testing-Issues and Applications. A training session presented at the Annual Meeting of the National Council on Measurement in Education.

Pollitt, A. (2012). The Method of Adaptive Comparative Judgement. *Assess. Educ. Principles, Pol. Pract.* 19, 281–300. doi:10.1080/0969594X.0962012.066535410.1080/0969594x.2012.665354

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danmarks Paedagogiske Institut.

Sijtsma, K. (2009). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika* 74, 107–120. doi:10.1007/s11336-008-9101-0

Stark, S., and Chernyshenko, O. S. (2011). Computerized Adaptive Testing with the Zinnes and Griggs Pairwise Preference Ideal point Model. *Int. J. Test.* 11, 231–247. doi:10.1080/15305058.2011.561459

Van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., and De Maeyer, S. (20162016). Validity of Comparative Judgement to Assess Academic Writing: Examining Implications of its Holistic Character and Building on a Shared Consensus. *Assess. Educ. Principles, Pol. Pract.* 26, 59–74. doi:10.1080/0969594X.2016.1253542

Van Daal, T. (2020). Making a Choice Is Not Easy?!: Unravelling The Task Difficulty of Comparative Judgement to Assess Student Work *[Doctoral Dissertation*. Antwerp, Belgium: University of Antwerp.

Verhavert, S., De Maeyer, S., Donche, V., and Coertjens, L. (2018). Scale Separation Reliability: What Does it Mean in the Context of Comparative Judgment? *Appl. Psychol. Meas.* 42, 428–445. doi:10.1177/0146621617748321

Verhavert, S., Bouwer, R., Donche, V., and De Maeyer, S. (2019). A Meta-Analysis on the Reliability of Comparative Judgement. *Assess. Educ. Principles, Pol. Pract.* 26, 541–562. doi:10.1080/0969594X.2019.1602027

Wright, B. D., and Stone, M. H. (1979). *Best Test Design*. Chicago, IL: MESA Press.

Conflict of Interest: EC was employed by the company Cito.

The remaining 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

MINI REVIEW
published: 03 March 2022
doi: 10.3389/feduc.2022.787926

A Review of the Valid Methodological Use of Adaptive Comparative Judgment in Technology Education Research

Jeffrey Buckley[1,2]*, *Niall Seery*[1] *and Richard Kimbell*[3]

[1] Faculty of Engineering and Informatics, Technological University of the Shannon: Midlands Midwest, Athlone, Ireland, [2] Department of Learning, KTH Royal Institute of Technology, Stockholm, Sweden, [3] Goldsmiths, University of London, London, United Kingdom

There is a continuing rise in studies examining the impact that adaptive comparative judgment (ACJ) can have on practice in technology education. This appears to stem from ACJ being seen to offer a solution to the difficulties faced in the assessment of designerly activity which is prominent in contemporary technology education internationally. Central research questions to date have focused on whether ACJ was feasible, reliable, and offered broad educational merit. With exploratory evidence indicating this to be the case, there is now a need to progress this research agenda in a more systematic fashion. To support this, a critical review of how ACJ has been used and studied in prior work was conducted. The findings are presented thematically and suggest the existence of internal validity threats in prior research, the need for a theoretical framework and the consideration of falsifiability, and the need to justify and make transparent methodological and analytical procedures. Research questions now of pertinent importance are presented, and it is envisioned that the observations made through this review will support the design of future inquiry.

Keywords: comparative judgment, technology education, design, validity, methodology, assessment

OPEN ACCESS

Edited by:
Renske Bouwer,
Utrecht University, Netherlands

Reviewed by:
Christian Bokhove,
University of Southampton,
United Kingdom
Wei Shin Leong,
Ministry of Education, Singapore
Scott Bartholomew,
Brigham Young University,
United States

Correspondence:
Jeffrey Buckley
jeffrey.buckley@tus.ie;
jbuckley@kth.se

Specialty section:
This article was submitted to
Assessment, Testing and Applied
Measurement,
a section of the journal
Frontiers in Education

Received: *01 October 2021*
Accepted: *24 January 2022*
Published: *03 March 2022*

Citation:
Buckley J, Seery N and Kimbell R
(2022) A Review of the Valid
Methodological Use of Adaptive
Comparative Judgment in Technology
Education Research.
Front. Educ. 7:787926.
doi: 10.3389/feduc.2022.787926

INTRODUCTION

Technology education is relatively new to national curricula at primary and secondary levels in comparison to subjects such as mathematics, the natural sciences, and modern and classic languages. Broadly, technology education relates to subjects focused on thinking and teaching about technology (de Vries, 2016), with subjects taking different formats internationally (cf., Buckley et al., 2020b). For example, in Ireland there are four technology subjects at lower secondary level and four at upper secondary level. In contrast, in England the single subject of Design and Technology is offered at Key Stages 1, 2, and 3 of secondary education. A central feature of contemporary technology education is an emphasis on "nurturing the designerly" (Stables, 2008; Milne, 2013). Design tasks are therefore prominent within the technology classroom, the outcome of which is usually a portfolio of work and accompanying artifact which evidence the process and product of learning. While these portfolios, in response to the same activity, can vary widely in length, content, and content type, it would be typical to see progression from initial sketches and notes representing "hazy ideas," through stages of idea refinement, to the technical presentation of a final proposed solution (e.g., Kimbell et al., 2009; Seery et al., 2012).

With pedagogical approaches in technology education growing in empirical support (cf., McLain, 2018, 2021), the integration of design has been problematized from the perspective of constructive alignment (Buckley et al., 2020b). A critical challenge remains in how, given the variety of ways through which technology learners can demonstrate capability (Kimbell, 2011), such as through varied portfolios, educators can validly and reliably assess open-ended, designerly outputs, without imposing an assessment architecture which infringes on the validity and meaningfulness of the associated learning processes. Comparative judgment (CJ), particularly adaptive comparative judgment (ACJ), is presented within the pertinent literature as auspicious in that it would appear to solve this particular disciplinary problem. The process of ACJ is described in detail by Hartell and Buckley (2021), but in brief it involves a cohort of assessors, typically referred to as "judges," who individually make holistic pairwise comparisons on digital or digitized representations of student work which are subjected to assessment, i.e., portfolios (Kimbell, 2012; Pollitt, 2012a,b). Over a series of rounds, judges make value-laden, binary judgments on portfolios which are selected for comparison based on an adaptive sorting algorithm (Canty, 2012). Ultimately, this results in a rank order from "best" to "worst" with relative differences presented as parameter values. The attributes which lend to ACJ being a solution to the assessment of designerly outputs are that the rank order is derived through a consensus of the judging cohort which has been proven to be highly reliable, and it mitigates issues with traditional criterion referenced assessment stemming from rubrics which can lack content validity and which are difficult to implement reliably (Sadler, 2009).

Research on the use of ACJ in technology education is rising continuously (Bartholomew and Jones, 2021). However, the research questions which are investigated tend to be broad and relate to whether ACJ is feasible and whether it is appropriate and reliable in the assessment of designerly outputs. The resounding answer to these questions is "yes." ACJ has been shown to be highly reliable in each relevant study which presents reliability statistics (Kimbell, 2012; Bartholomew and Yoshikawa-Ruesch, 2018; Bartholomew and Jones, 2021) and its validity can be seen as tied to the assessors (Buckley et al., 2020a; Hartell and Buckley, 2021) with outputted misfit statistics being useful to audit or gain insight into outlying judges or portfolios (Canty, 2012). While many of the conducted studies have taken the form of mechanistic, efficacy and effectiveness studies through the use of correlational and experimental designs, the research has largely been exploratory due to the lack of a theoretical framing regarding the place of ACJ within the technology classroom. Further, while in this research ACJ is examined as an assessment instrument, it is used as a research instrument in the collection of original data. This overlap in purpose has resulted in noteworthy limitations and validity threats as ACJ is a complex system which makes it difficult to interpret specific study results as the reason for any improved education outcomes. Given that ACJ can be used to assess designerly learning, and that the existing exploratory evidence indicates educational benefit, there is now a need to progress this research agenda in a more rigorous and systematic fashion.

With a view toward advancing this agenda, this article presents a review of existing ACJ studies relating to technology education. The intent of which is to highlight aspects of this area of scholarship which require methodological refinement to guide the design of future studies and to pose critical research questions stemming from existing evidence which are of immediate importance. This is of particular significance to technology education as ACJ has developed technologically to the point where it is becoming more frequently adopted in research and practice for both formative (Dewit et al., 2021) and summative purposes (Newhouse, 2014). Further, the agenda to "evolve" the use of ACJ for national assessment in technology education has been laid out by Kimbell (2012), and if this is to be successful the underpinning evidence base needs to be robust.

Two useful systematized reviews have already been conducted by Bartholomew and Yoshikawa-Ruesch (2018) and Bartholomew and Jones (2021) with aims of consolidating the pertinent evidence. Using the search outcomes of these two reviews, a combined total of 38 articles (see **Supplementary Table 1** for details), a qualitative review and synthesis is herein conducted of which the outcomes are presented thematically in the following sections. Unlike the prior reviews which have been valuable in summarizing the outcomes of ACJ investigations, this paper presents a critical review of limitations in how ACJ has been investigated (cf., Grant and Booth, 2009). A critical review does not necessarily include a systematic search process, although the articles reviewed here result from two (Bartholomew and Yoshikawa-Ruesch, 2018; Bartholomew and Jones, 2021). The intent of a critical review is to "take stock" of the value of prior contributions through critique. Critical reviews do not intend to provide solutions, but rather questions and guidance which may "provide a "launch pad" for a new phase of conceptual development and subsequent "testing" (Grant and Booth, 2009, p. 93). The review process included a thorough review of each sampled article in terms of the alignment and appropriateness of presented aims and/or research questions, methodological approaches, data analysis, and conclusions drawn. Any limitations identified were then conceptually grouped into "themes" through a process of pattern coding (Saldaña, 2013). The themes are presented, not with an exhaustive critique of each reviewed article, but as summaries with descriptions and exemplars.

THEMES RELATING TO AREAS FOR IMPROVEMENT IN ACJ SCHOLARSHIP IN TECHNOLOGY EDUCATION RESEARCH

Theme 1: Validity Threats Through Making Inference Beyond What the Generated Evidence Can Support

The validity of ACJ as an assessment instrument is frequently commented on. What is often not discussed is the validity of the use of ACJ in research studies and associated validity threats. Due to the ethical implications of randomized control trials in denying students access to what researchers believe to be impactful for their learning (De Maeyer, 2021), much ACJ research in technology education is quasi-experimental. Inferences from this research, however, are often made which such a methodology cannot support. To take one example, Bartholomew et al. (2019a) present a quasi-experiment where at the mid-way point of a design project, each student in an experimental group made 17 judgments using ACJ on their peers work, where a control group engaged with a peer-sharing activity reflective of traditional practice. At the end of the study, all portfolios were combined into a single ACJ assessment session, but only the teacher and experimental group students acted as judges. The authors observed a significant difference in that the experimental group on average outperformed the control group and concluded that "our analysis suggests that students who participate in ACJ in the midst of a design assignment reach significantly better levels of achievement than students who do not" (p. 375). However, the inference that ACJ could be causal is not supported. The effect, for example, could have come from the experimental group simply being exposed at the mid-way point to a greater volume of examples (an exposure effect), to having to make judgments on quality or critique peer work (a judgment effect), or as only the experimental group assessed all work at the end, they may have judged in favor of familiar work (a recognition effect). Subsequent work addressed many of these limitations by mitigating the possible recognition effect (Bartholomew et al., 2020a), and the on-going "Learning by Evaluating" project (Bartholomew and Mentzer, 2021) is actively pursuing the qualification of explicit effects which can stem from ACJ, a need commented on further in Theme 2. A related issue comes from Newhouse (2014) where a cohort of judges noted that the digitized work presented in the ACJ session was a poor representation of the actual student work. One assessor commented on how the poor quality of some photographs made it more difficult to see faults which were easier to see in real life. This comment raises an important issue which is not regularly commented on—the use of ACJ may be valid from a process perspective, but if the portfolios are not accurate representations of the students learning or capability itself, the outcome of the ACJ session may be invalid. Through the review there were multiple examples where authors made inferences or suggestions which they could not support based on the described study. This is not to say that the studies themselves had no value or contribution—they have—but it is important not to infer beyond what an implemented methodology can substantiate.

Theme 2: Theoretical Framing to Define the Many Elements of Adaptive Comparative Judgment

Extending on the previous theme, nearly all studies where ACJ was used as an intervention which reported a positive effect attributed the effect to ACJ as a whole. In these studies, ACJ is often used by students in a way to support their learning (e.g., Bartholomew et al., 2019a; Seery et al., 2019). There is a need to move beyond this broad inference. The use of ACJ could offer educational benefit when learners act as judges through exposure to the work of peers, having to critique and compare the quality of work, having to explicate comments justifying a decision, or a combination of the these. The research needs to move to a stage of identifying the activity which has the educational benefit if it is to make a more significant contribution to knowledge. Further, all these activities can be conducted without an ACJ system in a classroom. Educators could organize activities where learners are exposed to, compare, and constructively critique the work of their peers outside of an ACJ software solution. The pedagogical benefits of the activities inherent to ACJ could be more easily transferred to classrooms if the focus of ACJ research was on defining the important processes rather than the broad benefit of the system holistically when used for learning.

The need to investigate the nuances of ACJ makes the need of a theoretical framework for ACJ apparent, and this would need to consider the intended purpose of ACJ, i.e., assessment as, for, or of learning. Related concepts merit further definition, in particular "time" and "criteria." Many studies examine the efficiency of ACJ in comparison to traditional assessment practices (Rowsome et al., 2013; Bartholomew et al., 2018a, 2020b; Zhang, 2019) however, for ACJ time is usually considered in terms of total or average judging time. There is need to consider any set-up or training times to give a truer reflection of the impact this could have on practice, and any comparisons would need to consider the time educators put into developing rubrics and repeat usage as well. Similarly, many studies aim to determine judging criteria (Rowsome et al., 2013; Buckley et al., 2020a) but to understand the implications of such work, a theoretical framework which identifies whether criteria are relevant at a topic level, task level, or as specific as an individual judgment level merits qualification.

Theme 3: Validity in the Determination of Validity

The need for a theoretical framework for ACJ also encompasses the need to determine how claims can be falsified. Given the strength of evidence illustrating that ACJ is reliable, many efforts have turned to the valid use of ACJ. Specifically, the question is presented as to whether ACJ is a more valid alternative to traditional criterion reference assessment in the assessment of designerly student work. The validity of the rank can be assumed if (1) the cohort of judges is determined as appropriate, i.e., the rank is a valid representation of their consensus, and (2) judgments are based on reasoned decisions, i.e., judges take the task seriously and there are no technical errors (Buckley et al., 2020a). The first assumption is a decision of judge selection. For the second, Canty (2012) describes how misfit statistics can be used to identify outlier judges who importantly could have made reasoned judgments but are outliers in terms of having a different view of capability or learning than the majority of the cohort. Multiple studies use correlations between an ACJ rank and grades generated through the use of traditional rubrics as a measure of validity (Canty, 2012; Seery et al., 2012; Bartholomew et al., 2018a,b, 2019b; Strimel et al., 2021). Based on these studies, while not explicit, an implicit suggestion is being made that the hypothesis that ACJ offers a valid measure of assessment could be falsified if non-significant or negative correlations were observed in these investigations. If the study begins with a critique of rubrics, the issue is that the validity of ACJ is being determined by how closely it can re-produce the grades of the tool it is presented as being the better alternative to (e.g., Seery et al., 2012). This is further compounded by concerns regarding the content validity of rubrics for the assessment of design learning and who the assessors are. For example, the correlation between an ACJ and traditional rubric generated ranks when both are generated by experts has a very different meaning than if one rank comes from students. If the used rubrics are not critiqued in this way and are determined as valid, this application is not necessarily problematic.

Theme 4: There Is a Need to Justify Approaches to Statistical Data Analysis

A pedagogically useful attribute of ACJ stems from the parameter values within the final rank of portfolios. These follow a cubic function (Kimbell et al., 2007; Kimbell, 2012) and offer insight into relative performance between portfolios. This is commonly noted as a significant benefit of ACJ (Bartholomew et al., 2020b; Buckley et al., 2020a) and its potential was demonstrated by Seery et al. (2019) where parameter values were transposed into student grades. However, despite articles claiming the benefit of parameter values over the rank order which is linear and thus does not present relative differences, much of the data analysis does not utilize these values. Importantly, it may not be appropriate to use parameter values if model assumptions for parametric tests are violated. However, none of the reviewed articles which presented a formal statistical analysis provided any details of model assumptions which were tested. Statistical tests used have been both parametric and non-parametric, but this selection appears random. Where non-parametric tests are used it may be that authors are choosing to adopt tests which do not require certain assumptions to be met and which are more robust to outliers, but such a reason is not provided. Further, there was evidence of important information such as test statistics and/or degrees of freedom not being reported (e.g., Bartholomew et al., 2019b, p. 13) and only statistically significant results being reported with a note that there were non-significant results which were not presented (e.g., Bartholomew et al., 2017, p. 10). This is common in technology education research more generally (Buckley et al., 2021b), and is suggestive of the need for further transparency in data analysis.

Theme 5: Transparency in Adaptive Comparative Judgment Research

A final theme, which extends on occasional missing information in reported statistical tests relates more broadly to levels of transparency in the reporting of ACJ studies. There is a general need to improve levels of transparency in technology education research (Buckley et al., 2021a) and it was notable, particularly in conference publications that the methodology sections were not comprehensive enough for readers to fully understand the nature of investigations (e.g., Canty et al., 2017, 2019). The information which tended to be omitted was details on the design tasks that students would have engaged with, of which outcomes were assessed through ACJ. It is probable that this relates to space limitations with conference papers and that the authors would be providing this information during the conference presentation, but it would be useful to provide such information as an appendix, perhaps through an open access repository if space limitations are the issue. Finally, making research transparent relates not just to describing in detail how a study was conducted, but also to providing rationales for decisions which are made (Closa, 2021). No study which was conducted offered a clear justification of sample size. Study populations and sampling procedures were explained, but authors, to date, have not considered either empirical of ethical implications of having samples sizes which are too small or excessively large. It would be appropriate if, as this research progresses, decision making around sampling is made more apparent.

DISCUSSION

Research using and on the use of ACJ in technology education to date has been useful in demonstrating that student work which is generated through the ill-defined and open-ended activities reflective of contemporary technology education can be reliably assessed. It is also clear that the validity of ACJ can be qualified in many ways, such as through the careful design of the judging cohort and by making use of misfit statistics. ACJ has been repeatedly observed as capable of providing reliable ranks and positive educational effects when used for learning, and the research to date has identified many important considerations such as that portfolios need to be accurate representations of the objects of assessment. Due to how often these outcomes have been observed, it is questionable whether further inquiry into these broad research questions would lead to any further insight. Instead, as an outcome of this review it is recommended that ACJ research becomes more systematic, nuanced, and explicit. Foremost, there is a need for appropriately designed methodologies and caution needs to be given when making inferential claims, but there are also ethical considerations associated with investing further resources into studies examining outcomes which have been repeatedly observed. For example, ACJ is continuously observed to be reliable, however, no studies have been conducted which examine a core proposition of this—that the reliability stems from the aggregation of judgments from cohorts of assessors with individual biases. It would be useful to examine the reliability of ACJ when the judging cohort is purposefully selected to include people with differing opinions, or who are provided with different criteria to make judgments on, in attempts to falsify this claim. Further, on this point and extending on the need for a theoretical framework outlined in theme 2, there is need to consider how reliable ACJ needs to be depending on its intended use, e.g., summative vs. formative, and what are the associated educational implications of different reliability thresholds (cf., Benton and Gallacher, 2018).

This need for more systematic inquiry creates the need for ACJ researchers to develop a theoretical framework. A current question is not whether the use of ACJ when used for learning (typically involving students as judges) has educational merit, but why could and why has ACJ been observed to have a positive effect? It is paramount that central concepts such as time/efficiency and criteria are adequately defined, and recognition must be given that at present it can be difficult for teachers to use ACJ due to, for example, cost and training implications. However, the nature of activity within the ACJ process such as making comparative judgments or being exposed to large variation in student work is immediately accessible to teachers as pedagogical approaches. There is significant potential for research to be conducted, either using or not using ACJ, which provides insight into the value of ACJ and which is immediately transferable into practice. The next phase of ACJ research should focus less on broad questions of feasibility and potential holistic benefit, and instead focus more on refining the use of ACJ for practice and on identifying the components of the ACJ process which have positive effects on learning and the student experience.

AUTHOR'S NOTE

This is a critical review, which included a self-review of the authors own published works.

AUTHOR CONTRIBUTIONS

JB conceptualized the study and wrote the first draft. All authors then reviewed and edited the manuscript.

FUNDING

The authors received no explicit financial support for the conduction of this research. The publication of this article was supported by the Faculty of Engineering and Informatics at the Technological University of the Shannon: Midlands Midwest, Ireland.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2022.787926/full#supplementary-material

REFERENCES

Bartholomew, S., and Jones, M. (2021). A systematized review of research with adaptive comparative judgment (ACJ) in higher education. *Int. J. Technol. Des. Educ.* 1–32. doi: 10.1007/s10798-020-09642-6

Bartholomew, S., and Mentzer, N. (2021). *Learning by Evaluating: Engaging Students in Evaluation as a Pedagogical Strategy to Improve Design Thinking. Community for Advancing Discovery Research in Education*. Available online at: https://cadrek12.org/projects/learning-evaluating-engaging-students-evaluation-pedagogical-strategy-improve-design (accessed November 17, 2021).

Bartholomew, S., Mentzer, N., Jones, M., Sherman, D., and Baniya, S. (2020a). Learning by evaluating (LbE) through adaptive comparative judgment. *Int. J. Technol. Des. Educ.* doi: 10.1007/s10798-020-09639-1

Bartholomew, S., Reeve, E., Veon, R., Goodridge, W., Lee, V., and Nadelson, L. (2017). Relationships between access to mobile devices, student self-directed learning, and achievement. *J. Technol. Educ.* 29, 2–24. doi: 10.21061/jte.v29i1.a.1

Bartholomew, S., Strimel, G., and Jackson, A. (2018a). A comparison of traditional and adaptive comparative judgment assessment techniques for freshmen engineering design projects. *Int. J. Eng. Educ.* 34, 20–33.

Bartholomew, S., Strimel, G., and Yoshikawa, E. (2019a). Using adaptive comparative judgment for student formative feedback and learning during a middle school design project. *Int. J. Technol. Des. Educ.* 29, 363–385. doi: 10.1007/s10798-018-9442-7

Bartholomew, S., Strimel, G., and Zhang, L. (2018b). Examining the potential of adaptive comparative judgment for elementary STEM design assessment. *J. Technol. Stud.* 44, 58–75. doi: 10.2307/26730731

Bartholomew, S., Yoshikawa, E., Hartell, E., and Strimel, G. (2020b). Identifying design values across countries through adaptive comparative judgment. *Int. J. Technol. Des. Educ.* 30, 321–347. doi: 10.1007/s10798-019-09506-8

Bartholomew, S., and Yoshikawa-Ruesch, E. (2018). "A systematic review of research around adaptive comparative judgement (ACJ) in K-16 education," in *CTETE - Research Monograph Series*, ed. J. Wells (Virginia: Council on Technology and Engineering Teacher Education), 6–28. doi: 10.21061/ctete-rms.v1.c.1

Bartholomew, S., Zhang, L., Bravo, E. G., and Strimel, G. (2019b). A tool for formative assessment and learning in a graphics design course: adaptive comparative judgement. *Des. J.* 22, 73–95. doi: 10.1080/14606925.2018.1560876

Benton, T., and Gallacher, T. (2018). Is comparative judgement just a quick form of multiple marking? *Res. Matters* 26, 22–28.

Buckley, J., Adams, L., Aribilola, I., Arshad, I., Azeem, M., Bracken, L., et al. (2021a). An assessment of the transparency of contemporary technology education research employing interview-based methodologies. *Int. J. Technol. Des. Educ.* doi: 10.1007/s10798-021-09695-1

Buckley, J., Canty, D., and Seery, N. (2020a). An exploration into the criteria used in assessing design activities with adaptive comparative judgment in technology education. *Irish Educ. Stud.* doi: 10.1080/03323315.2020.1814838

Buckley, J., Hyland, T., and Seery, N. (2021b). Examining the replicability of contemporary technology education research. *Tech. Series* 28, 1–9. doi: 10.1016/j.jgg.2018.07.009

Buckley, J., Seery, N., Gumaelius, L., Canty, D., Doyle, A., and Pears, A. (2020b). Framing the constructive alignment of design within technology subjects in general education. *Int. J. Techn. and Des. Educ.* 31, 867–883. doi: 10.1007/s10798-020-09585-y

Canty, D. (2012). *The Impact of Holistic Assessment Using Adaptive Comparative Judgement on Student Learning*. Doctoral Thesis. Limerick: University of Limerick.

Canty, D., Buckley, J., and Seery, N. (2019). "Inducting ITE students in assessment practices through the use of comparative judgment," in *Proceedings of the 37th International Pupils' Attitudes Towards Technology Conference*, eds S. Pule and M. de Vries (Msida, Malta: PATT), 117–124.

Canty, D., Seery, N., Hartell, E., and Doyle, A. (2017). *Integrating Peer Assessment in Technology Education Through Adaptive Comparative Judgment*. Philadelphia: Millersville University, 1–8.

Closa, C. (2021). Planning, implementing and reporting: increasing transparency, replicability and credibility in qualitative political science research. *Eur. Political Sci.* 20, 270–280. doi: 10.1057/s41304-020-00299-2

De Maeyer, S. (2021). *Reproducible Stats in Education Sciences: Time to Switch? Reproducible Stats in Education Sciences*. Available online at: https://svendemaeyer.netlify.app/posts/2021-03-24_Time-to-Switch/ (accessed April 26, 2021).

de Vries, M. (2016). *Teaching About Technology: an Introduction to the Philosophy of Technology for Non-philosophers*. Switzerland: Springer.

Dewit, I., Rohaert, S., and Corradi, D. (2021). How can comparative judgement become an effective means toward providing clear formative feedback to students to improve their learning process during their product-service-system design project? *Des. Technol. Educ.* 26, 276–293.

Grant, M. J., and Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Info. Libr. J.* 26, 91–108. doi: 10.1111/j.1471-1842.2009.00848.x

Hartell, E., and Buckley, J. (2021). "Comparative judgement: An overview," in *Handbook for Online Learning Contexts: Digital, Mobile and Open*, eds A. Marcus Quinn and T. Hourigan (Switzerland: Springer International Publishing), 289–307.

Kimbell, R. (2011). Wrong but right enough. *Des. Technol. Educ.* 16, 6–7.

Kimbell, R. (2012). Evolving project e-scape for national assessment. *Int. J. Technol. Des. Educ.* 22, 135–155. doi: 10.1007/s10798-011-9190-4

Kimbell, R., Wheeler, T., Miller, S., and Pollitt, A. (2007). *E-scape Portfolio Assessment: Phase 2 Report*. London: Goldsmiths, University of London.

Kimbell, R., Wheeler, T., Stables, K., Shepard, T., Martin, F., Davies, D., et al. (2009). *E-scape Portfolio Assessment: Phase 3 Report*. London: Goldsmiths, University of London.

McLain, M. (2018). Emerging perspectives on "the demonstration" as a signature pedagogy in design and technology education. *Int. J. Technol. Des. Educ.* 28, 985–1000. doi: 10.1007/s10798-017-9425-0

McLain, M. (2021). Developing perspectives on 'the demonstration' as a signature pedagogy in design and technology education. *Int. J. Technol. Des. Educ.* 31, 3–26. doi: 10.1007/s10798-019-09545-1

Milne, L. (2013). Nurturing the designerly thinking and design capabilities of five-year-olds: technology in the new entrant classroom. *Int. J. Technol. Des. Educ.* 23, 349–360. doi: 10.1007/s10798-011-9182-4

Newhouse, C. P. (2014). Using digital representations of practical production work for summative assessment. *Assess. Educ.* 21, 205–220. doi: 10.1080/0969594X.2013.868341

Pollitt, A. (2012b). The method of adaptive comparative judgement. *Assess. Educ.* 19, 281–300. doi: 10.1080/0969594X.2012.665354

Pollitt, B. (2012a). Comparative judgement for assessment. *Int. J. Technol. Des. Educ.* 22, 157–170. doi: 10.1007/s10798-011-9189-x

Rowsome, P., Seery, N., Lane, D., and Gordon, S. (2013). "The development of pre-service design educator's capacity to make professional judgments on design capability using adaptive comparative judgment," in *Paper Presented at 2013 ASEE Annual Conference & Exposition Proceedings*, (Atlanta: ASEE).

Sadler, D. R. (2009). "Transforming holistic assessment and grading into a vehicle for complex learning," in *Assessment, Learning and Judgement in Higher Education*, ed. G. Joughin (Netherlands: Springer), 45–63.

Saldaña, J. (2013). *The Coding Manual for Qualitative Researchers*, 2nd Edn. Los Angeles: SAGE.

Seery, N., Buckley, J., Delahunty, T., and Canty, D. (2019). Integrating learners into the assessment process using adaptive comparative judgement with an ipsative approach to identifying competence based gains relative to student ability levels. *Int. J. Technol. Des. Educ.* 29, 701–715. doi: 10.1007/s10798-018-9468-x

Seery, N., Canty, D., and Phelan, P. (2012). The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. *Int. J. Technol. Des. Educ.* 22, 205–226. doi: 10.1007/s10798-011-9194-0

Stables, K. (2008). Designing matters; designing minds: the importance of nurturing the designerly in young people. *Des. Technol. Educ.* 13, 8–18.

Strimel, G. J., Bartholomew, S. R., Purzer, S., Zhang, L., and Ruesch, E. Y. (2021). Informing engineering design through adaptive comparative judgment. *European J. Eng. Educ.* 46, 227–246. doi: 10.1080/03043797.2020.1718614

Zhang, L. (2019). *Investigating Differences in Formative Critiquing Between Instructors and Students in Graphic Design*. West Lafayette: Purdue University.

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Education

PERSPECTIVE
published: 18 March 2022
doi: 10.3389/feduc.2022.806894

Using Teachers' Judgments of Quality to Establish Performance Standards in Technology Education Across Schools, Communities, and Nations

Niall Seery[1]*, *Richard Kimbell*[2] *and Jeffrey Buckley*[1,3]

[1] *Faculty of Engineering and Informatics, Technological University of the Shannon, Athlone, Ireland,* [2] *Goldsmiths, University of London, London, United Kingdom,* [3] *Department of Learning, KTH Royal Institute of Technology, Stockholm, Sweden*

The establishment and maintenance of national examination standards remains a serious issue for teachers and learners, whilst the levers of control remain firmly in the hands of Awarding Bodies and supervising politicians. Significantly, holistic assessment presents an agility and collective approach to establishing in the minds of teachers "what is of value" when determining the comparative evidence of pupil performance. It is argued in this paper that the collation of the comparative judgment process can initially identify and subsequently maintain standards of performance that can be defined on a cluster, regional or even national level. Much comparative judgment research centers on the formative benefits for learners, but here we place the focus on teachers operating in collaborative groups to establish standards within and beyond their own schools, and ultimately across the nation. We model a proof-of-concept research project. A rank is produced by the collective consensus of the participating teachers and used to simulate a definition of standard. Extrapolations are statistically modeled to demonstrate the potential for this approach to establishing a robust definition of national standards. But central to the process is what is going on in the minds of teachers as they make their judgements of quality. The research aims to draw out teachers' constructs of quality; to make them explicit; to share them across classrooms and schools; and to empower teachers to debate and agree their standards across schools. This research brings to the fore the symbiotic relationship between teaching, learning and assessment.

Keywords: national standards, teacher judgment, comparative judgment, validity, assessment

Edited by:
Tine Van Daal,
University of Antwerp, Belgium

Reviewed by:
Wei Shin Leong,
Ministry of Education, Singapore

****Correspondence:***
Niall Seery
nseery@ait.ie

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *01 November 2021*
Accepted: *02 February 2022*
Published: *18 March 2022*

Citation:
Seery N, Kimbell R and Buckley J (2022) Using Teachers' Judgments of Quality to Establish Performance Standards in Technology Education Across Schools, Communities, and Nations. Front. Educ. 7:806894. doi: 10.3389/feduc.2022.806894

INTRODUCTION

Much of the focus of Adaptive Comparative Judgment (ACJ) research is centered on cohort-based application cases (Williams and Kimbell, 2012; Bartholomew and Jones, 2021), where the agendas include formative (e.g., Bartholomew et al., 2019) and summative (e.g., Jones and Alcock, 2012; Whitehouse and Pollitt, 2012) application, sometimes combining both formative and summative agendas to frame an assessment "as" learning approach (Seery et al., 2012). Studies report high levels of reliability (Bartholomew and Yoshikawa-Ruesch, 2018; Bartholomew and Jones, 2021), and this gives confidence in the rank order produced by the binary judging session. ACJ uses an adaptive algorithm to govern the presentation of pairs of student "portfolios" which are then holistically compared by a cohort of "judges," e.g., teachers, on evidence of learning. Research by the Assessment of Performance Unit (APU) team at Goldsmiths in the 1980s empirically demonstrated that teachers were far more reliable when comparatively assessing whole pieces of work than they were when assessing individual qualities. When assessing writing performance through a comparative approach, teacher judgment is reported as "highly internally consistent when judging quality" (Heldsinger and Humphry, 2010, p. 221). These binary decisions on authentic evidence ultimately position students' work on a rank from top to bottom, as described by Pollitt (2012a,b). The approach produces associated parameter values or "ability scores" which are indications of relative differences between portfolios along the rank.

Considering the context of Design and Technology, the creative relationship between designing and making is difficult to reflect in a criterion driven assessment and this approach can in fact change the very nature of the activity to conform to what is weighted as valued output. The challenge in developing a retrospective portfolio and even an artifact is influenced by the criterion specified before the task begins. The work of the Kimbell et al. (1991) established the need to consider student work differently and framed the importance of a holistic view of performance. Like consensual assessment techniques (CAT; Amabile, 1982), the aggregation of expert judgements through an ACJ approach provides a reliable and valid approach to measuring (Bramley, 2015; Bramley and Wheadon, 2015; Coertjens et al., 2017; Verhavert et al., 2018; Kimbell, 2021). Aligned with the approach of using calibrated exemplars (Heldsinger and Humphry, 2013), this research proposes a "bottom up" development of national standards.

RESEARCH AGENDA

Up to this point, ACJ has been used with groups of learners with the purpose of arriving at a performance rank of those learners for formative and/or summative purposes. If 100 learners are participating in an ACJ session, then, using ACJ, judges can arrive at a performance rank for those 100 learners. But the practice becomes more complex if large cohorts are anticipated, as would be the case for national examinations. Annually in Ireland, approximately 60,000 students take examinations as part of the Leaving Certificate—a State organized national examination taken at the end of secondary level education with results feeding into a matriculation system for tertiary education admissions. Managing such a number through an ACJ exercise would be extremely challenging. However, it is not the purpose of this project to attempt such an exercise. Rather we seek to investigate the use of a new form of ACJ to begin to explore what would be involved in building a system that enables teachers to collaborate across schools to arrive at a view of a national performance standard. Broadly speaking this system would start with a locally established standard (within a school or small cluster of schools) and move progressively to regional groupings of schools (e.g., across cities/counties) and ultimately to a view of a standard across the entire nation.

To begin this inquiry, three initial steps are required to first establish a performance standard at a local level:

1. First, standards will be tentatively defined and agreed between teachers in collaborating schools in terms of learners' performance on an ecologically valid activity.
2. Second, the work will be combined into a single ACJ session which will be judged by the teachers who had supervised the work in the collaborating schools. This is a classroom-based view of standards in which teachers do not seek to apply a standard devised by someone else, but rather they will create a rank of the work using holistic judgments guided by their own personal constructs of capability which can then be used to refine and clarify standards.
3. Third, through a collaborative process, teachers will reflect upon the ACJ process and refine the initially determined standards and then map these back onto the rank order produced through the ACJ session to create a "reference scale" of work which other, new pieces of work can be positioned along.

This research agenda is to establish a reference scale that can be used to position students' work relative to a national standard, built from an initial local but representative standard, and in essence determine national standards through authentic performance. It is useful to think of this "Steady State" as a type of *Ruler*. Importantly, the Ruler would be produced from authentic evidence of student work, brought about by the judgments of teachers making decisions on authentic work. This Ruler reference would move away from abstracted criteria or the need for interpretation, instead the real evidence of learners' work would form the basis of comparators, representing the quality of work presented. The idea is that this approach would improve standards by improving the whole performance and supporting the developed conceptualization of what constitutes quality.

Central to the establishment of the Ruler is drawing from teachers "what is of value" when judging student work. This research centers on using the holistic approach of ACJ to establish a reference rank that is built on teacher expertise. Jones et al. (2015) argue that freeing judges from predefined criteria can enable judges to tap into their expertise. This view is supported by van Daal et al. (2019) and is seen as particularly useful when assessing complex competencies (Pollitt, 2012a). Comparative methods enable the teacher to obtain reliable scores for complex skills more easily than using analytic methods (Jones and Inglis, 2015; Coertjens et al., 2017). Barkaoui (2011) also found that holistic marking favored higher levels of consistency between markers in comparison with analytic marking. The work of Lesterhuis et al. (2018) is also relevant as their findings report that the considered construct of quality in a comparative method is multidimensional and notes that teachers use their experience to give meaning to the evidence. Not only can the "bottom up" approach to standards setting unleash the expertise of teachers, but building on the insights shared by Van Gasse et al. (2019) where they highlight the change in the conception of the assessor following a CJ intervention, it can also support the development of the teacher in terms of refining their constructs of capability. The richness of varying perspectives discussing the concrete context of the assignment, ensures task-orientated and not examiner-orientated focus. As highlighted by Heldsinger and Humphry (2010), the potential of using a calibrated sale of exemplars to compare future work is the technical focus of this proposed work.

MODELING AN APPROACH TO DEFINING NATIONAL STANDARDS

The planned scoping project will operate through three phases, and the subject of study will be "Design and Communication Graphics" which has an annual Higher Level examination cohort of circa 4,000 students. To get a representative sample to define a national sample, the following approach and calculations have been produced.

- A total population of 3,804 is the subject of this proof-of-concept research cohort and a sample of 250 portfolios was calculated to give a confidence level of 95% with a margin of error of 5.99%.
- A selection of judges will participate in a normal ACJ session with the work. Judging sub-groups will be identified to include teachers, researchers, and experienced examiners. The representative rank order will be produced by the judges making binary decisions on a combined pool of portfolios from all over the country, so the teachers are not merely judging their own learners but are exposed to a representative sample of evidence. This will produce an agreed performance rank for the 250 learners based on the judgments of the teachers, researchers, and examiners. Reliability statistics will be carefully examined to explore any differential effects of the teachers' judgments. The performance rank will thereafter be called The Ruler.
- An additional sample of 50 portfolios will also be randomly selected from the population data that are independent from the work that created the Ruler. At the end of the project, the resulting 50 pieces of work will be judged against the Ruler using a modified ACJ tool. The purpose is to (1) explore approaches to positioning these along the Ruler and then (2) to see where and how the 50 new pieces locate themselves along the Ruler. The judgments will be made by the original team of teachers, researchers, and examiners.
- The research questions will focus on the length and precision of the Ruler by exploring teacher judgment and the variations in the ways things can be valued. The process of consensus building on what is of value within the process of building the ruler and ensuring the bias management and representation test the lens appropriate to its utility as an instrument for national standard definition.

The agenda to establish a means by which a national standard can be determined from the evidence produced by pupils and adjudicated on by in-practice teachers, highlights several research considerations discussed in the following section.

GAP ANALYSIS

Extending the application of the ACJ process beyond that of a single cohort or cluster to form a national picture of performance brings into focus the details of the ACJ reliability statistics, parameter values and the association with validity (cf. Buckley et al., 2022) all of which are of particular interest to this research. Although reported reliability statistics of more than 0.9 give confidence (Pollitt, 2012b; Bartholomew and Jones, 2021), there are notable critiques of the adaptive process. Bramley has identified that the adaptive algorithm artificially inflates estimates of the reliability of the outputted rank order of work (Bramley, 2015; Bramley and Wheadon, 2015). Much of the issue is caused by a "spurious separation among the scripts" (Bramley, 2015, p. 14) where the adaptive algorithm makes it impossible, for example, for work that "loses" a small number (e.g., two) of judgments against work when paired truly at random to show that it is actually relatively good work as the adaptive algorithm will make it less likely to get paired with work that won in those initial rounds. Further, the process of ACJ has issues at the extremes. To take the piece of work that "wins" or is ultimately placed at the top of the rank, it is likely that it may never or will rarely ever be judged as a losing piece of work in a pairwise comparison. As such, there is little information about the work compared to those determined as closer to average. The winning piece of work may confidently be positioned at the top, but there is much uncertainty regarding its parameter value. While these issues do not affect the absolute rank, they do affect the validity of interpreting and using the parameter values as denotations of relative distances between pieces of work, which is problematic when we seek to develop the application beyond a single cohort.

A number of studies have aimed to address these problems. First, Bramley and Vitello (2019) note some potential advantages of adaptivity. These included that adaptivity can increase efficiency by avoiding pairing portfolios which are very far apart on the rank, and on the issue of inflated reliability they note that while adaptivity may spuriously inflate the standard deviation, it could actually reduce error. One possible approach to addressing inflated reliability which will be explored in this project is to increase the number of comparisons. Verhavert et al. (2019) note that in CJ, to reach a reliability of 0.90 without adaptivity, 26–37 comparisons are needed per portfolio. Bramley and Vitello (2019) point out that the reason for the inflated SD is that the introduction of adaptivity means most portfolios would be compared indirectly via other portfolios. Therefore, it is possible that the use of adaptivity to select the portfolios for comparison which would provide the most information with a minimum number of comparisons, such as 26–37 per portfolio, used as a stopping rule as opposed to the use of a reliability threshold could provide a suitable solution. Such a minimum number will need to

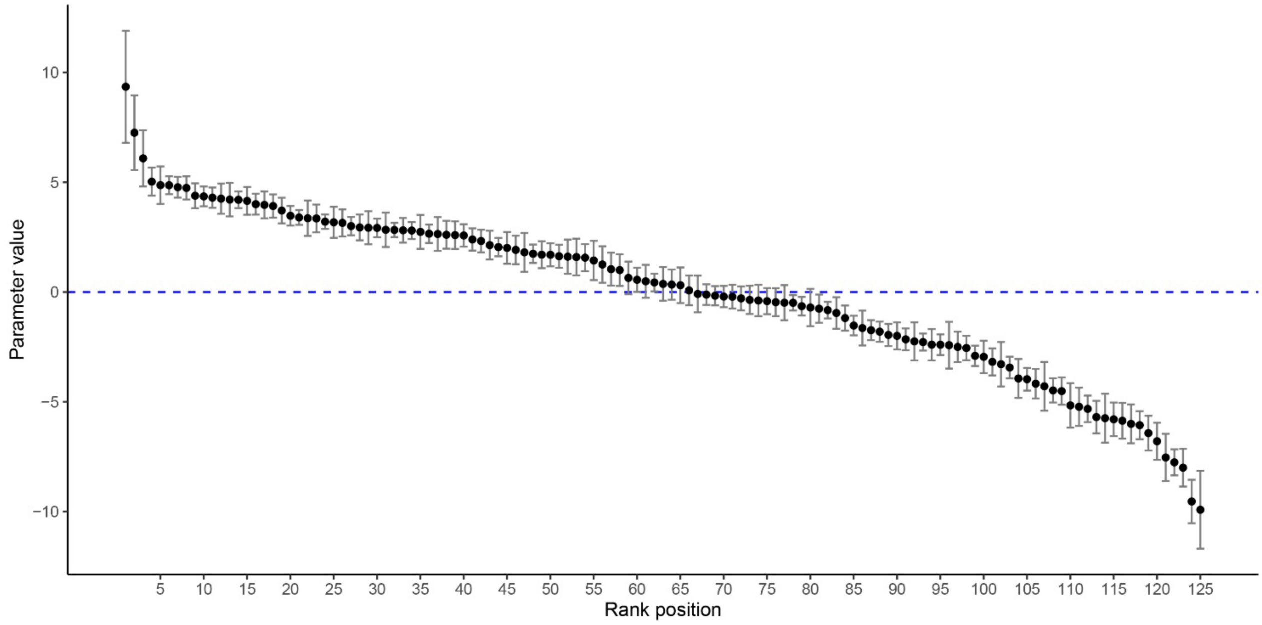


FIGURE 1 | The goal of step 1 is to develop a normative rank order of work, which includes information regarding locally derived standards inductively generated from reflection on authentic work. Note that the above figure consists of simulated data from an arbitrarily defined 125 pieces of work for illustrative purposes only.

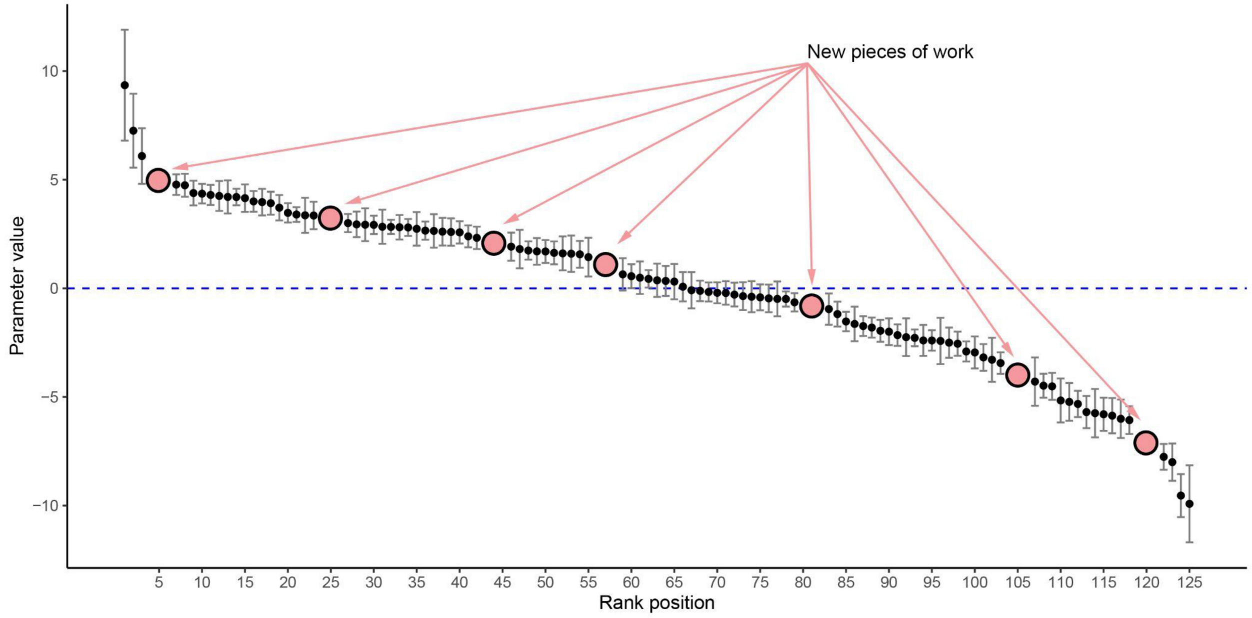


FIGURE 2 | The goal of step 2 is to use the Ruler, with a new ACJ approach, to position new pieces of work against, illustrated hypothetically in the above figure.

be determined, and in doing so reliability deflation (Bramley and Vitello, 2019) should be considered.

The research therefore will involve a two-stage process. First creating the Ruler (through a normal ACJ judgment process—**Figure 1**) and second employing the Ruler in a separate judgment process to seek to locate new pieces of work within the quality scale defined by the Ruler (**Figure 2**). This second process clearly requires a different ACJ approach. In "normal" ACJ judging sessions, all the work is floating and is affected by each judgment. A new comparison judgment will therefore affect the position of both pieces of work involved in the comparison. What we are proposing is that once the Ruler has been agreed it is fixed, and judgments thereafter (in the 2nd phase of judging) are intended simply to locate the new pieces within the Ruler.

The standards articulation process can occur at several points. It can be inserted at the end of the first judging round and lead to an articulation by teachers of the Ruler, but then it can occur again after the second phase of judging locates the new work into the Ruler. Understanding these standards is at the heart of this project and teachers will become very familiar with iterative discussions that seek to clarify and refine them. It is our belief that teachers have a working understanding of quality standards that enable them to distinguish good from mediocre work and mediocre from poor performance. But these standards are typically internal to their practice. Our aim is to draw them out through a process that (1) requires the teachers to use them in a judging round, and (2) through discussion empowers teachers to articulate what indications and qualities they see in the work that makes them judge it as outstanding/good/adequate/poor. Whitehouse and Pollitt (2012, p. 15) highlight that "thought needs to be given to how shared criteria can be exemplified and disseminated." At the end of the second phase a range of statistical exercises can be undertaken with the resulting data. One might judge, for example, that the distribution of the new pieces of work from the second pair of schools was loaded more toward the upper end of the Ruler. This can be calculated exactly. The Ruler does not merely show individual placings but can also reveal school-based performance.

This process can go on as often as required with school groups of work being endlessly judged against the Ruler, producing individual positions for the work and school–based data from the amalgamation of those placings. All schools in a region (and even across the nation) can therefore be assessed with an ACJ-judgment-style of assessment against a common standard—the Ruler. This does not require an enormous "once-for-all" ACJ exercise simultaneously involving thousands of learners. Rather it can be done in two steps by (1) establishing the Ruler and (2) subsequently comparing learners from other schools to the Ruler. The concept of a ruler affords the utility of an instrument that can order authentic work on a scale representative of the breath of performance. The disparate parameter values record the separation between units of work, enabling transposition of the rank normatively or relatively, depending on the sensitivity of the assessment context. Like previous research, ACJ can also record the judge's statistical alignment, unpacking further consensus and misfit.

Building on comparative judgment, teacher assessment and professional development in various subject contexts, this paper proposes a study that will endeavor to answer questions that focus on 3 thematic areas: considerations for teachers and schools, technological developments, and standards and awards, with the details being unpacked in the following sections.

Teachers and Schools

- Can teachers use authentic data (with ACJ judging) to articulate what standards are?
- Can teachers fully articulate what distinguishes "good" from "less good" work?
- Can teachers' decisions reach consensus and be aggregated so as to determine what is of value when considering evidence of learning?

Technological Developments

- Can we establish a valid and reliable definition of standards and thereby create a Ruler that is long enough and precise enough to cater for all performance levels?
- From a technological perspective, how can "new" work be judged into the Ruler?

Awards and Standards

- Can teachers distinguish statistically discrete levels of performance from within the Ruler?
- Can teachers use the Ruler to effectively compare other work to a National Standard?

Supplementary research questions that will be explored in parallel and not as part of the modeling study include:

- What is the impact of exposing teachers to a breadth of work from other schools on their definition of standards?
- How will this exposure impact teachers' professional development, specifically in assessment literacy?
- What is the relationship between task design and student performance?
- Do teachers' articulation of standards vary with the task?
- Are there inherent biases that impact on different categories of students?
- Can assessment tasks be designed to be independent? Or can we control task independence?
- Can we (or should we) articulate national standards as absolute and monitor performance over time?
- How could the Ruler be used in practice as a formative and pedagogical tool?

CRITICAL ISSUES FOR TEACHERS AND SCHOOLS

The first and most critical feature of this approach to standard-setting is that the standards emerge directly because of the judgments made by teachers. Whilst teachers' ACJ judgments will be informed by criteria, those criteria are not individually scored and summed. Rather, they are all "held-in-mind" to support the teacher in making an overall holistic judgment of the quality of the work. Teachers' concepts of quality, Polanyi (1958) referred to this quality as connoisseurship, are central to the approach we seek to build. Teachers discuss the strengths and weaknesses of individual pieces in a comparison, and the many finely distinguished pieces in the Ruler provide a scale that exemplifies quality at every level. Wiliam (1998) described teachers doing this in the early days of the England/Wales National Curriculum. Given pages of criteria to score, they largely ignored them, preferring to make their judgments in more holistic ways:

> . . . most summative assessments were interpreted not with respect to criteria (which are ambiguous). . . but rather by reference

> to a shared construct of quality that exists in well-defined communities of practice. (p. 6).

The work of Seery et al. (2012), exemplify the capacity of ACJ to help build quality constructs, using actual evidence as the medium for refining the emerging constructs of novice student teachers.

> A critical factor for the students was that the assessor (their peer) could empathize with their work having completed the process themselves. The process also encouraged students to engage in discussions on capability with their peers in an effort to broaden their concept and understanding of capability as the ACJ model sees judgments on students' work made across a wide range of assessors. (p. 224).

It is these constructs of quality that we shall be exploring within the community of graphic teachers. The aim will be both to build and enrich these explicit constructs and, in the process, to enable teachers to see their own learners' work against a wider frame of reference than exists in their own school. With another school . . . in another town . . . and ultimately across the nation. As teachers become more familiar with the quality of work that can be expected in relation to any task, they are empowered to develop their own practice and help their learners to improve.

CRITICAL ISSUES FOR THE "RULER"

Refining ACJ to accommodate the national standards agenda will support several critical agendas. The judging process will engage teachers in developing an understanding of what standards are, this is especially powerful when these standards are being defined by actual classroom-based evidence, where differentiation between work is established by qualities adjudicated by the teachers (van Daal et al., 2019). Building a dataset that can represent and define national standards has the capacity to build confidence in standards across schools, a bi-directional relationship where the feed-forward micro to macro definition of standards will also backwash from macro to micro to help teachers to improve the performance of their own students.

The significance of the national standard definition is critically dependent on the quality of the Ruler. Therefore, the Ruler needs to be long enough (so it captures the full range of performance, with no loss of utility at the ends) and precise enough (so work can be accurately placed on the Ruler). The statistical and technological solutions to developing a robust Ruler are apparent challenges. More nuanced are the challenges facing teachers in determining to what degree work can be distinguished into distinct units of performance and how many distinct units of performance can be measured at each grade level. Statistically, this plays out in terms of the standard deviation of the items (portfolios) and the discrimination that can be achieved by the teachers, both are critical to the reliability that can be achieved in the judging (Kimbell, 2021).

The Ruler can only be as good as the work that is imported into the algorithm. We could have a very good Ruler for the average piece of work, but a poor Ruler for excellent work. This is a key research agenda. Assuming the target will be a normal distribution, the focus of the research agenda is to ensure the precision and length of the Ruler, to cater for the full spectrum of performance. There are several approaches that could be used to test the robustness of the Ruler. We could bias the population sample or chain the judging session to "force" judging of comparisons within specific areas of the rank (at the ends for example, where usually we have the least amount of information). Using the analogy of a Microscope, the technology could be designed to have interchangeable lenses to take a focused look at categories of interest not just performance bands, but also (for example) issues of inclusion, access, and disadvantage. This perspective and approach have not ever been made manifest in earlier or even current ACJ work and is only necessary when you consider using the rank for the purposes, we intend here, for inter-school, clustered or national standard definition.

There are 3 critical issues to be considered. The ACJ algorithm and its ability to refine the information captured in relation to the spread of performance, requires critical and statistical review. That is, the length of the Ruler and the resultant graduations are sufficiently defined to represent the breadth of performance that then can be used for future comparisons. Secondly, the probability at the extremes needs to be comparable with the confidence in parameter values that emerge in the middle of the rank, with no risk of inflating the reliability of the rank. The criticism of inflation at the extremes, needs to be managed to ensure that the graduations of the analogous Ruler are consistent at the extremes and can distinguish performance effectively. Thirdly, once the Ruler is created and robustly tested the issue is to translate or transpose the rank order into a definition of standards. The creation of standards will rely on the experience of the teachers in distinguishing the discrete units of performance to form a robust dataset that represents the breadth or performance and can confidently identify grade boundaries.

CRITICAL ISSUES FOR EXAMINATION AND AWARDING BODIES

Perhaps the most critical issue for examination bodies in this approach exists in the question "How do we create the Ruler?" It would be simple enough to take a sample of schools and use that sample to create it with (say) 100 or 500 learners. But how do we ensure that it is sufficiently broadly based to capture all the levels of quality that we are concerned to identify? One possibility would be to see the Ruler as emergent and evolving, based on the standards of last year's examinations, and enriched with this year's work samples. It might therefore contain some of last year's work samples as well as this year's. This would additionally provide the possibility of a direct comparison of standards across years.

And this raises the question of the variability of performance across tasks. Examinations do not set the same questions every year. But, since they are looking to assess the same qualities, the assumption is that different questions can elicit parallel levels of performance. With task-based performance in graphics it will be interesting to see how (to what extent) parallel levels of performance can be revealed by different tasks. And critical to that will be the articulation of the standards themselves. Teachers will initially seek to clarify their standards in relation to the task-based performance of the initiating group. But the articulation process must be sufficiently generic that it applies beyond the detail of the task itself.

There is a fine line here. The standards emerge from task-based performance, since it is the task-based performance that exemplifies those standards for the teachers to observe. But the standard needs to operate equally on parallel tasks, so it must be sufficiently specific to operate accurately on a given task but still sufficiently generic to accommodate variation. The details of this inter-task dynamic will be very revealing of teachers' views of the standard.

The paper is the starting point of a comprehensive research study that sets out to develop existing technology that has the potential to liberate teachers' professional judgment through engagement with authentic evidence of pupil learning, while establishing a definition of national standards.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

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

FUNDING

The authors received no explicit financial support for the preparation of this manuscript. Costs to support the publication of this manuscript were provided by RM Education.

REFERENCES

Amabile, T. M. (1982). Social psychology of creativity: a consensual assessment technique. *J. Pers. Soc. Psychol.* 43, 997–1013. doi: 10.1037/0022-3514.43.5.997

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assess. Educ. Princ. Policy Pract.* 18, 279–293. doi: 10.1080/0969594X.2010.526585

Bartholomew, S., and Jones, M. (2021). A systematized review of research with adaptive comparative judgment (ACJ) in higher education. *Int. J. Technol. Des. Educ*. doi: 10.1007/s10798-020-09642-6

Bartholomew, S., Strimel, G., and Yoshikawa, E. (2019). Using adaptive comparative judgment for student formative feedback and learning during a middle school design project. *Int. J. Technol. Des. Educ*. 29, 363–385. doi: 10.1007/s10798-018-9442-7

Bartholomew, S., and Yoshikawa-Ruesch, E. (2018). "A systematic review of research around adaptive comparative judgement (ACJ) in K-16 education," in *CTETE - Research Monograph Series*, ed. J. Wells (Reston, VA: Council on Technology and Engineering Teacher Education), 6–28. doi: 10.21061/ctete-rms.v1.c.1

Bramley, T. (2015). *Investigating the Reliability of Adaptive Comparative Judgment*. Cambridge: Cambridge Assessment.

Bramley, T., and Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assess. Educ. Princ. Policy Pract.* 26, 43–58. doi: 10.1080/0969594X.2017.1418734

Bramley, T., and Wheadon, C. (2015). "The reliability of adaptive comparative judgment," in *Paper Presented at the AEA–Europe Annual Conference*, Glasgow, 7–9.

Buckley, J., Seery, N., and Kimbell, R. (2022). A review of the valid methodological use of adaptive comparative judgement in technology education research. *Front. Educ.* 6. doi: 10.3389/feduc.2022.787926

Coertjens, L., Lesterhuis, M., Verhavert, S., Van Gasse, R., and De Maeyer, S. (2017). Judging texts with rubrics and comparative judgement: taking into account reliability and time investment. *Pedagog. Stud.* 94, 283–303.

Heldsinger, S. A., and Humphry, S. M. (2013). Using calibrated exemplars in the teacher-assessment of writing: an empirical study. *Educ. Res.* 55, 219–235. doi: 10.1080/00131881.2013.825159

Heldsinger, S., and Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *Aust. Educ. Res.* 37, 1–19. doi: 10.1007/BF03216919

Jones, I., and Alcock, L. (2012). "Summative peer assessment of undergraduate calculus using adaptive comparative judgement," in *Mapping University Mathematics Assessment Practices*, eds P. Iannone and A. Simpson (Norwich: University of East Anglia), 63–74.

Jones, I., and Inglis, M. (2015). The problem of assessing problem solving: can comparative judgement help? *Educ. Stud. Math.* 89, 337–355. doi: 10.1007/s10649-015-9607-1

Jones, I., Swan, M., and Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *Int. J. Sci. Math. Educ.* 13, 151–177. doi: 10.1007/s10763-013-9497-6

Kimbell, R. (2021). Examining the reliability of adaptive comparative judgement (ACJ) as an assessment tool in educational settings. *Int. J. Technol. Des. Educ.* doi: 10.1007/s10798-021-09654-w

Kimbell, R., Stables, K., Wheeler, A., Wosniak, A., and Kelly, V. (1991). *The Assessment of Performance in Design and Technology*. London: Schools Examinations and Assessment Council/Central Office of Information.

Lesterhuis, M., Van Daal, T., Van Gasse, R., Coertjens, L., Donche, V., and De Maeyer, S. (2018). When teachers compare argumentative texts: decisions informed by multiple complex aspects of text quality. *L1 Educ. Stud. Lang. Lit.* 18, 1–22. doi: 10.17239/L1ESLL-2018.18.01.02

Polanyi, M. (1958). *Personal Knowledge: Towards a Post–Critical Philosophy*. Chicago, IL: University of Chicago Press.

Pollitt, A. (2012a). Comparative judgement for assessment. *Int. J. Technol. Des. Educ.* 22, 157–170. doi: 10.1007/s10798-011-9189-x

Pollitt, A. (2012b). The method of adaptive comparative judgement. *Assess. Educ. Princ. Policy Pract.* 19, 281–300. doi: 10.1080/0969594X.2012.665354

Seery, N., Canty, D., and Phelan, P. (2012). The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. *Int. J. Technol. Des. Educ.* 22, 205–226. doi: 10.1007/s10798-011-9194-0

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., and de Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assess. Educ. Princ. Policy Pract.* 26, 59–74. doi: 10.1080/0969594X.2016.1253542

Van Gasse, R., Lesterhuis, M., Verhavert, S., Bouwer, R., Vanhoof, J., Van Petegem, P., et al. (2019). Encouraging professional learning communities to increase the shared consensus in writing assessments: the added value of comparative judgement. *J. Prof. Cap. Commun.* 4, 269–285.

Verhavert, S., Bouwer, R., Donche, V., and De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assess. Educ. Princ. Policy Pract.* 26, 541–562. doi: 10.1080/0969594X.2019.1602027

Verhavert, S., De Maeyer, S., Donche, V., and Coertjens, L. (2018). Scale separation reliability: what does it mean in the context of comparative judgment? *Appl. Psychol. Meas.* 42, 428–445. doi: 10.1177/0146621617748321

Whitehouse, C., and Pollitt, A. (2012). *Using Adaptive Comparative Judgement to Obtain a Highly Reliable Rank Order in Summative Assessment*. Manchester: Centre for Education Research and Policy.

Wiliam, D. (1998). "The validity of teachers' assessments," in *Paper Presented to Working Group 6 (Research on the Psychology of Mathematics Teacher Development) of the 22nd Annual Conference of the International Group for the Psychology of Mathematics Education*, Stellenbosch, 1–10.

Williams, P. J., and Kimbell, R. (2012). Special issue on e-scape [Special issue]. *Int. J. Technol. Des. Educ.* 22, 123–270.

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 27 April 2022
doi: 10.3389/feduc.2022.802392

Exploring the Validity of Comparative Judgement: Do Judges Attend to Construct-Irrelevant Features?

Lucy Chambers and Euan Cunningham*

Research Division, Cambridge University Press & Assessment, Cambridge, United Kingdom

When completing a comparative judgment (CJ) exercise, judges are asked to make holistic decisions about the quality of the work they are comparing. A key consideration is the validity of expert judgements. This article details a study where an aspect of validity, whether or not judges are attending to construct-irrelevant features, was investigated. There are a number of potentially construct-irrelevant features indicated in the assessment literature, and we focused on four features: appearance; handwriting; spelling, punctuation, and grammar (SPaG); and missing response vs. incorrect answer. This study explored this through an empirical experiment supplemented by judge observation and survey. The study was conducted within an awarding organisation. The particular context was within a programme of work trialling, a new method of maintaining examination standards involving the comparative judgement of candidates' examination responses from the same subject from two different years. Judgements in this context are cognitively demanding, and there is a possibility that judges may attend to superficial features of the responses they are comparing. It is, therefore, important to understand how CJ decisions are made and what they are or are not based on so that we can have confidence in judgements and know that any use of them is valid.

Keywords: comparative judgment, standard maintaining, construct-irrelevance, validity, assessment

Edited by:
Marije Lesterhuis,
Spaarne Gasthuis, Netherlands

Reviewed by:
Niall Seery,
Athlone Institute of Technology,
Ireland
Ben Davies,
University College London,
United Kingdom

****Correspondence:***
Lucy Chambers
Lucy.Chambers@Cambridge.org

Specialty section:
This article was submitted to
Assessment, Testing and Applied
Measurement,
a section of the journal
Frontiers in Education

Received: *26 October 2021*
Accepted: *14 March 2022*
Published: *27 April 2022*

Citation:
Chambers L and Cunningham E (2022) Exploring the Validity of Comparative Judgement: Do Judges Attend to Construct-Irrelevant Features? Front. Educ. 7:802392. doi: 10.3389/feduc.2022.802392

INTRODUCTION

The study was conducted within an English awarding organisation, where each year thousands of candidates' examination scripts[1] are scrutinised by trained experts. We often think of *marking* as the primary activity within this context; however, there are other routine activities that involve a *holistic assessment* of scripts, namely standard setting (deciding on a cut-score for a grade boundary) and standard maintaining (ensuring the chosen cut-score represents the same standard as previous years). Recently, a programme of work exploring an alternative method for standard maintaining was conducted that used comparative judgment (CJ) of candidates' examination scripts (henceforth scripts). In the process of trialling this method, a key consideration is the validity of expert judgements. This article details a study where an aspect of validity, whether or not judges are attending to construct-irrelevant features, was investigated. An evaluation of the method itself is beyond the scope of this study and is presented in Benton et al. (2020b, 2022).

[1] Examination script is the term used to denote a candidate's question responses contained in answer sheets or an answer booklet.

In a framework for evidencing assessment validity developed by Shaw et al. (2012), one of the central validation questions is "Are the scores/grades dependable measures of the intended constructs?" (p.167). It follows that, for scores to be valid, judgements informing them must also be based on the intended constructs. The emphasis on intended constructs noted here is key for CJ; judges should base their decisions on construct-relevant features and avoid any influence of construct-irrelevant features (Messick, 1989). For example, in the assessment context, judgements influenced by an appropriate use of terminology would be construct-relevant, whereas those based on the neatness of handwriting would not be. CJ is a technique whereby a series of paired or ranked judgements (typically made by multiple judges) is used to generate a measurement scale of artefact quality (Bramley, 2007; Pollitt, 2012a,b). For example, pairs of candidate scripts can be compared in order to judge which script in each pair is the "better" one or packs of scripts can be ranked in order from best to worst. Analysis of these judgements generates an overall rank order of artefacts, in this case, scripts, and a scale of script quality (in logits) is created with each script having a value on this scale. One of the main advantages of CJ is that it requires judges to make relative judgements, which are sometimes considered to be easier to make than absolute judgements, e.g., of an individual script against a mark scheme (Pollitt and Crisp, 2004).

When completing a CJ exercise, judges are asked to make holistic decisions about the quality of the work they are comparing. Judges are not given specific features to focus on; instead, they draw on their experience to make the judgements. In an assessment context, this open holistic nature of the decision is very different from that of a traditional marking decision, which often follows a strict mark scheme. This difference is exacerbated if the judgement increases from an item-based decision to one based on an entire script.

When making holistic decisions, judges can decide what constitutes good quality; in practice, this conceptualisation can vary across judges. If judges are attending to construct-irrelevant features, then this could have implications for validity. In addition, as each script is viewed by multiple judges, the final rank order is determined by the combined decision-making of multiple judges. If judges' conceptualisations do not cover every relevant dimension of the construct, then this again has implications for validity (van Daal et al., 2019). Thus, the validity of CJ is comprised of both the individual holistic nature of decision-making and the fact that the final rank order is based on a shared consensus or the collective expertise of judges (van Daal et al., 2019). A focus on construct-irrelevant features could impact both of these elements.

In a study investigating written conceptions of mathematical proof, Davies et al. (2021) explored which features judges collectively valued using CJ. One aspect of the study compared the CJ results of two groups of participants, the first comprised a group of expert mathematicians and the second comprised a group of educated non-mathematicians. This enabled divergent validity to be explored, i.e., judgements of the experts were based on mathematical expertise rather than on surface features such as grammar and quality of the writing. They found a modest correlation between the two sets of scores, and non-expert judges failed to produce a reliable scaled rank order for the writing samples. This study suggests that mathematical expertise was key to the task; however, it does not eliminate the possibility that attention was given to construct-irrelevant features.

Turning to assessment, "To date, not much is known about which aspects guide assessors' decisions when using comparative methods" (Lesterhuis et al., 2018, p.3). Previous research investigating the validity of CJ decision-making has mostly utilised decision statements (Whitehouse, 2012; Lesterhuis et al., 2018; van Daal et al., 2019), and to our knowledge, there is only one experimental study (Bramley, 2009). A discussion of these studies will follow.

Decision statements are post-decision judge reflections "explaining or justifying their choice for one text over the other" (Lesterhuis et al., 2018, p.5), and they help to shed light on the criteria judges use. In a study using decision statements to explore the validity of CJ decision-making in academic writing, van Daal et al. (2019) investigated whether there was full construct representation in the final rank order of essays. They found that, while the full construct was represented overall, representation did vary by judge. In addition, they found that additional construct-relevant dimensions were reported, suggesting that judges were drawing on their expertise. Lesterhuis et al. (2018) found that teachers considered wide ranging and multiple aspects of the text when investigating which aspects are important for teachers when making a CJ decision on argumentative texts. The teachers also paid great attention to more complex higher-order aspects of text quality. Interestingly, not all aspects were covered in each decision, suggesting some construct under-representation. The judges in this study also appeared to be utilising their experience. In a study involving teachers comparing geography essays, Whitehouse (2012) found that decision statements used the language contained in the assessment objectives and mark schemes. The judges would have been familiar with these mark schemes in their roles as teachers or examiners in the subject; Whitehouse speculated that this resulted in the creation of "their own shared construct" (p.12), which they used to make their decisions.

These three studies suggest that judges attended to multiple and varied construct-relevant aspects when making holistic decisions, and that they drew on their experience and shared construct. There are, however, limitations acknowledged by these authors in the use of specific research contexts and whether the method used fully elicited the entire range of aspects actually attended to. In addition, as with all self-report measures, there is a danger that judges may deliberately not report everything (e.g., as they know it is construct-irrelevant) or they may not know or be able to verbalise what they attended to.

Bramley (2009) attempted to circumnavigate these methodological issues by conducting a controlled experiment. He prepared different versions of chemistry scripts, where each pair of scripts differed with respect to only one potentially construct-irrelevant feature. In total, four features were manipulated across 40 pairs of scripts: (i) the quality of written English; (ii) the proportion of missing as opposed to incorrect responses; (iii) the profile of marks in terms of fit to the Rasch model; and (iv) the proportion of marks gained on the subset of questions testing "good chemistry." These were then ranked by judges as part of a CJ exercise. The CJ script quality measures of the two versions were then compared to assess whether the feature in question influenced judgements. The method was successful in identifying that the largest effects were obtained for the following features: (ii) scripts with missing responses were ranked lower on average than those with incorrect responses and (iv) scripts with a higher proportion of good chemistry items were ranked higher on average than those with a lower proportion.

THIS STUDY

This study seeks to build on previous research to further explore judge decision-making, specifically whether or not judges are attending to construct-irrelevant features when making their CJ decisions. We did this by conducting an empirical experiment supplemented by judge observation utilising a think-aloud procedure and a post-task survey. Thus, we combined the objectivity of an experimental study with the richness of judges' verbalisations and actions and the explicitness of their *post hoc* reflection. If it was found that judges do pay attention to construct-irrelevant features when making judgements, then this has implications for how we use the results of CJ judgement exercises in this and potentially other contexts.

Standard maintaining, the context for our study, is the process whereby grade boundaries are set such that standards are maintained from 1 year to the next. CJ can be used in standard maintaining to provide information comparing the holistic quality of scripts from a benchmark test (e.g., June 18) with the holistic quality of scripts from a target test (e.g., June 19). Standard maintaining generally involves experts who are senior or experienced examiners. While these experts are used to the concept of holistic judgements, the current method used in England uses it in conjunction with statistical evidence. Making CJ decisions in this context without reference to any statistical or mark data, therefore, will be a novel experience for judges.

The explicit standard maintaining context itself adds another layer of complexity or difficulty to CJ decision-making, in that, it involves scripts from two different years. Judges, therefore, have to make complex comparisons (i) involving two sets of questions and answers and (ii) factoring in potentially differing levels of demand. These comparisons are cognitively demanding; it is, therefore, important to understand how CJ decisions are made and what they are or are not based on so that we can have confidence in the judgements.

The experimental method employed in this study draws on that of Bramley (2009) although set in a standard maintaining context. For this study, we also chose four construct-irrelevant features to investigate; however, all our script modifications were unidirectional (e.g., we always removed text to create missing responses), and we used a mixed-methods design incorporating judge observation with a think-aloud procedure.

There are a number of potentially construct-irrelevant features indicated in the assessment literature that could have an impact on marking or judge's decision-making. The majority of the research is marking-based, and findings have been mixed, with results often dependent on the subject and research context. Modification of some of these features could legitimately lead to a change in mark or script quality measure (henceforth CJ measure) depending on the qualification. We restricted the choice of features to those which should not cause a legitimate change in mark/CJ measure in the qualification used in the study, i.e., these features were not assessed as part of the mark scheme. From these, a number of features were conflated into four categories for use in this study:

- Appearance: crossings out/writing outside the designated area/text insertions.
- Handwriting: the effort required for reading (word-processed scripts were not included).
- Spelling, punctuation, and grammar (SPaG)[2].
- Missing: missing response vs. incorrect answer.

Findings from marking research that considered appearance reported that crossings out or responses outside the designated area decreased marker agreement (Black et al., 2011). This was even found for relatively straightforward items; Black et al. (2011) hypothesise "that the additional cognitive load of, say, visually dismissing a crossing-out, is enough to interfere with even simple marking strategies such as matching and scanning and hence increase the demands of the marking task" (p.10). Crisp (2013), in a study of teachers marking assessment coursework, found that two participants reported that features such as presentation and messy work are sometimes noted, where "the latter was thought to give the impression that the student does not care about the work" (p.10). Thus, negative predisposition to a script, in addition to increased cognitive load, may play a role in marking. To our knowledge, appearance has not been explored specifically in CJ tasks; this study investigated whether this feature interferes negatively with the complex demands of the CJ standard maintaining task.

The marking research findings around handwriting have been mixed, in varying contexts, and with few recent studies. Previous studies, described in Meadows and Billington (2005), have found that good handwriting attracted higher grades. This is perhaps because of the additional cognitive load involved in deciphering hard-to-read handwriting, e.g., it might take longer, cause frustration, or create doubt in the mind of the examiner. However, studies involving the United Kingdom examination boards with highly trained examiners and well-developed mark schemes have found no effect of handwriting on grades (Massey, 1983; Baird, 1998). In a second language testing context, Craig (2001) also found no influence of handwriting on test scores. In a study looking at the influence of script features on judgements in standard maintaining (not using CJ), paired comparisons, and rank ordering, Suto and Novakoviæ (2012) found that "no method was influenced to any great extent by handwriting" (p.17). It will be interesting to assess whether handwriting has an influence on highly trained examiners using an unfamiliar method of holistic comparative judgements as in this study.

[2] SPaG is part of the assessed construct for some qualifications but not for the qualification used for this study.

Spelling, punctuation, and, grammar (SPaG) has been found to influence student marks (Stewart and Grobe, 1979; Chase, 1983). For many qualifications, SPaG is part of the assessment construct; as a result, there has been limited recent research exploring any construct-irrelevant influence in a marking context. However, in a CJ context, Bramley (2009) found that manipulating SPaG in scripts had little influence on CJ measures. Also, in a CJ context, Curcin et al. (2019) found that SPaG was noted by judges, but, in comparison to subject-specific features, they were "considered little" (p.90). It will be beneficial to establish whether judges in this demanding and novel context study are influenced by SPaG.

In terms of missing response vs. incorrect answer, Bramley (2009) found that manipulating this feature in a controlled CJ experiment resulted in scripts with the missing responses being ranked lower on average than those with incorrect answers. Although not statistically significant (possibly because of a large SE), the size of the effect was approximately two marks. In a review of CJ and standard maintaining in an assessment context, Curcin et al. (2019) found that, in English language, missing responses "may have been used to some extent as 'quick' differentiators between scripts irrespective of the detailed aspects of performance" (p.89). Within both English language and literature, they found that missing responses influenced participant judgements "sometimes making them easier and sometimes more difficult" (p.94). Experimental modification of this feature will help us determine its effect on CJ standard maintaining decisions.

The results of modifying these four features in this experiment would provide evidence of whether certain construct-irrelevant variables are influencing the judging process. In addition to the CJ measures obtained through the experiment, we also collected information about which features judges were observed to attend to and which they reported attending to when making their judgements. This was obtained *via* a simplified think-aloud procedure and a questionnaire.

Our research question is given as follows: Are judges influenced by the following construct-irrelevant features when making CJ decisions in a standard maintaining context?

- Appearance.
- Handwriting.
- SPaG.
- Missing response vs. incorrect answer.

METHODS

Scripts

The study used a high-stakes school qualification typically sat at age 16 (GCSE). The examination was in Physical Education and was out of 60 marks. The format was a structured answer booklet that contained the questions and spaces for candidates to write their responses. There was a mixture of short answer and mid-length questions. This qualification was chosen because SPaG was not explicitly assessed. As the experiment was conducted in a standard maintaining context, it included scripts from both 2018 and 2019. As the features themselves are quite subjective, it was important for the researchers to establish a shared conceptualisation. Thus, before script selection took place, the researchers, in conjunction with the qualification manager, agreed definitions of the features (detailed in section "Features Defined").

For each year, 40 scripts were used, with one script on each mark point between 11 and 50. For 2018, these were randomly chosen. For 2019, ten scripts that exemplified each of the four features were chosen such that the marks were evenly distributed across the mark range (approximately one script in every five-mark block). **Figure 1** illustrates the scripts used in the study and how they relate to the starting scripts.

For the 2019 scripts, original and modified variants were needed. Modifications were made such that, if the modified scripts were re-marked in accordance with the qualification mark scheme, any changes should not result in an increase in mark. With the exception of the missing feature, the modified scripts were a positive variant of the feature in question, e.g., easier to read handwriting, improved SPaG, and neater appearance.

The researchers first detailed amendments that would be needed in the modified variants; for the SPaG and appearance features, these were checked by the qualification manager to ensure they were construct-irrelevant modifications. Forty volunteers were recruited to produce new variants of the 2019 scripts, with one volunteer per script. For SPaG, appearance, and missing features, both an original variant and a modified variant were made of each starting script. The original variant was a faithful reproduction of the starting script, just in the volunteer's handwriting. The modified variant was identical to the newly created original one apart from the specified modifications. This was to ensure that the only variable of change between the two variants was the feature in question. For the handwriting feature, only a new variant was produced. Again, this was a faithful

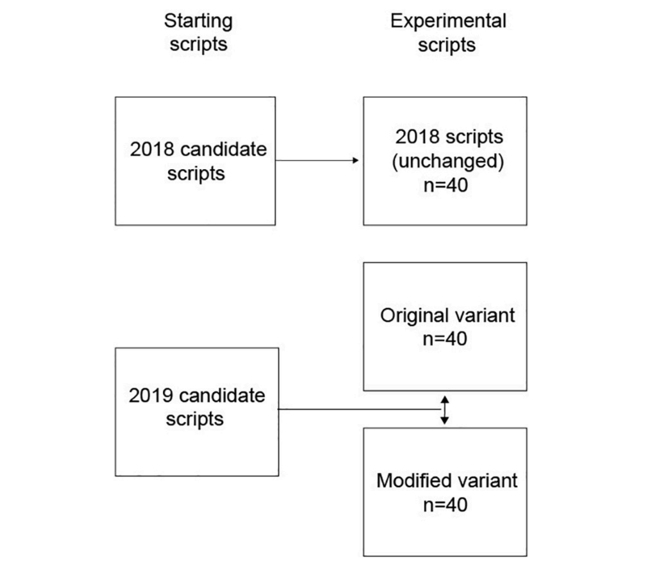


FIGURE 1 | Scripts used in the study.

reproduction of the starting script with no changes other than the handwriting. The researchers checked all the scripts to ensure that the conditions had been met.

Features Defined

Appearance

This feature included crossings out, text insertions, arrows pointing to another bit of text, and writing outside of the designated area. Examination rules for what is and is not marked were adhered to when making modifications. For example, for longer answers, an examiner would ignore any crossed-out text, so it could be removed in the modified variants; where there were text insertions or writing outside of the designated area, these were inserted into the main body of the text or the additional answer space as appropriate.

Handwriting

When defining problematic handwriting, we focused on the overall "effort" that was required to read a script. Thus, we chose scripts that were difficult to read; in practice, some of these scripts, at first glance, looked quite stylish. Writing that looked messy, or even just basic and very unsophisticated, but was easy to read was not included. When faced with a script that is hard to read, it can be hypothesised that an expert may award it a lower mark/rank, purely because the expert cannot establish whether it is correct, i.e., not the handwriting *per se*. Conversely, such a script may be given benefit of the doubt and get an appropriate or higher mark/rank. It should be noted that in traditional marking, examiners are asked to seek guidance from a senior examiner in cases where they are unable to read a response.

Spelling, Punctuation, and Grammar

Nearly all of the scripts contained some instances of non-standard grammar or punctuation. The scripts with non-standard SPaG tended to either contain many spelling errors, with reasonable punctuation and grammar, or the opposite. Scripts with non-standard spelling had errors in simple words or in words that were clearly taught on the course or that had even been used in the question that was being answered. For example, there were instances of the words "pulmonary" and "reversibility" being spelled in different ways within the same answer. Examples of non-standard grammar were the incorrect use of articles before nouns (e.g., "some gymnast," "these training programme"), the misuse of "they're," "their," and "there" and of "your" and "you're." Punctuation was generally lacking across many of the answers. Many of the scripts selected had limited punctuation. Examples included longer answers that were just one long sentence, apostrophes that were repeatedly used in the wrong place or not used at all, and full stops that were repeatedly used with no following capital letter. All modifications were made with reference to the mark scheme.

Missing

Scripts featuring a relatively high proportion of items that received zero marks but containing no more than two non-response answers were selected. Responses to some of these zero marked items were replaced with a non-response. This was based on the item omit rate calculated from the live examination and on plausibility (e.g., multiple choice answers and answers to the first few questions on the paper were not removed). As a result, these scripts had between six and fourteen non-responses largely depending on their total mark.

Judges

Ten judges were recruited from the examiner pool for the qualification; they were all experienced markers, and, in addition, two had experience of standard maintaining. They were either current or retired teachers of the course leading to the qualification. All the judges, therefore, had knowledge of the assessment objectives of the qualification, and through their marking experience, they would have gained a conceptualisation of what makes a good quality script. The judges were given information about CJ, standard maintaining, instructions on how to do the task, and information about the nature of the study. In order to re-familiarise themselves with the papers, they were given the two question papers and associated mark schemes. They were not presented with grade boundaries, but it should be noted that these are available publicly. The two papers used in this study were actually of a similar level of demand, i.e., had similar grade boundaries.

The decision on the number of judges used in the study was informed from an approximate power calculation based on the number of scripts, the fact that each script would be seen by each judge, and findings from previous CJ activities. The number of scripts used was based on balancing practicality (how many packs of scripts judges could feasibly judge alongside their work commitments, how many volunteers we could recruit to make the modifications, etc.) and sufficiency (having enough scripts to detect a difference).

Research Procedure

The original and modified variant 2019 scripts along with the 2018 scripts were presented to the judges embedded in a CJ standard maintaining exercise. The scripts were organised into packs of four, with each pack containing two 2018 scripts and two 2019 scripts (both original, both modified, or one of each). Packs of four were chosen, as the ranking of a script within a larger pack is more informative than whether it wins or loses a single paired comparison, so potentially, it is more efficient. Thus, in each pack, we had six comparisons rather than one (AB, AC, AD, BC, BD, and CD). The ordering of the four scripts within a pack was random: sometimes the first script in the list would be from 2018 and sometimes from 2019. Script allocation to each pack in terms of original marks was also random; thus, any pack could potentially contain scripts of similar or widely distributed original marks. The scripts and judging plan were loaded onto the in-house software used to conduct the experiment. In total, each judge would rank 20 packs, and they would see all the 2018 scripts but would only see either the modified or the original variant of each of the 2019 scripts.

Judges were presented with packs of four scripts and instructed to "rank these in order from best to worst overall performance." As the judges were all experienced examination markers of this qualification, they were asked to draw on this knowledge and experience and apply it to their CJ decisions.

No additional criteria beyond the mark scheme were provided, although the judges were given additional guidance on how to make holistic judgements. This included information on the importance of making an evaluation of the whole script and using their professional judgement to allow for differences in the questions and the relative difficulty of each test. The judges were aware that we were exploring a new method of conducting standard maintaining and were looking at how they made judgements, but they were unaware of the script modifications. The judges were informed of the script modifications and presented with a summary of the research findings at the end of the study.

The lead author observed each judge for approximately 30 min while they were making their judgements. This observation was conducted on Microsoft Teams, at a time of the judge's choosing; thus, it could be at the beginning, middle, or end of the judging period. The meeting software allowed the judges to share their screen, thereby allowing the observer to see what they were doing at any given point. This was supplemented by a think-aloud procedure in which the judges verbalised their thoughts while making their judgements. The judges were given the prompt "As you do the CJ task, we would like you to talk aloud about your actions, thoughts, and intentions. Please say anything that comes into your head while doing the task." To familiarise the judges with thinking aloud, they were given a short practice exercise (counting the number of windows in their house). The observation was recorded with the software, and this produced an automated transcript.

Once the judges had completed their judging, they were invited to complete a short online questionnaire. This gave the judges the opportunity to provide feedback and enabled us to gather additional information on their judging behaviour. In the questionnaire, we specifically asked the judges how they made their decisions.

Analysis

A mixed-methods design was used, which comprised a quantitative element derived from the CJ decision data and a qualitative element derived from the observation and survey responses.

We were interested in judge behaviour and, thus, wanted to check the quality and consistency of the judging. For this, we used the CJ decision data to calculate judge fit statistics, "judge fit is determined with regard to how well their judgements agree with what would be expected given the CJ measures of each script derived from the Bradley–Terry model" (Benton et al., 2020a, p. 10). This method does not use script marks. Typically, fit statistics are examined with a view to assessing whether any judges were misfitting the model to such an extent that they might be affecting judges' CJ decisions on the estimates of script quality. In some contexts, this might be a reason to exclude their judgements; but here, we were actually interested in the judges' behaviour, so no judges were removed on the basis of their fit statistics. Although the CJ data was collected as ranks, they were converted into pairs for judge fit analysis (A beats B, A beats C, B beats C, etc.). The fit analysis was completed using the Bradley Terry model (Bradley and Terry, 1952), and standard CJ fit statistics, infit and outfit mean-square statistics, were calculated in R (Wright and Masters, 1990; Linacre, 2002).

The main focus of the quantitative analysis was to establish whether the modified and original variants were judged to be of similar quality. The ranked CJ decision data, collected with the CJ tool, were analysed[3] using the Plackett-Luce model (Plackett, 1975). CJ measures were produced; these were based on which other scripts any given script were judged to be better or worse than and were calculated across multiple comparisons. These measures are logit values and are calculated for each script, indicating where a script sits on a constructed scale, which, in this case, was a measure of overall performance. As we were interested in whether the original and modified variants would be judged as being of similar quality, we compared the measures of the two variants. This was conducted by performing a paired t-test, which was calculated for each of the four features. Any significant results from the t-tests would indicate that the judges were attending to a particular construct-irrelevant feature when making their judgements. It should be noted that we treated the estimated CJ measures as error-free values (as we usually do with marks) in order to calculate t-tests; for this reason, their standard errors (SEs) were not utilised. Effect size was calculated using Cohen's D. Using the slope of regression lines calculated from comparing original marks to CJ measures, an approximate conversion factor of 1 logit equaling 5 marks was used to interpret effect sizes (after Bramley, 2009).

The qualitative element comprised of judge observation and survey. Each of the 10 judges was observed while performing their judging for approximately 30 min. While the verbalisations provide an indication of features being attended to, these features may not necessarily affect the actual decision-making. However, the analysis of the observation data does provide additional context with which to interpret the empirical analyses. It is possible that the behaviour exhibited during the observation did not reflect the rest of the judging; however, given the candid comments made by the judges, the authors suggest that it is unlikely to have been fundamentally different.

The script recordings and auto-generated transcripts of the judge's observations were loaded into qualitative analysis software. First, parts of the transcripts where the judges spoke about their decision-making or features they attended to were cleaned and corrected. Then, a targeted thematic analysis was conducted that involved coding across the four experimental features and other potentially construct-irrelevant features. As this was a simple coding exercise, looking at the presence or absence of mentions of the four features and any other potential construct-irrelevant feature, we involved only one researcher in the analysis, and no inter-rater coding reliability exercise was carried out. In order to maximise the accuracy of the data, the coding was completed in two stages; (1) when viewing the full recordings and (2) on a separate occasion through keyword analysis of the transcripts (using the text analysis tools available in the software). Responses to the post-task questionnaire were analysed along similar themes. When reporting the findings, all quotes are written in italics; those from

[3] This can be done using the R package *Plackett-Luce* (Turner et al., 2020).

the observations are written verbatim, and for those from the survey responses, spelling was corrected and punctuation was added to improve readability.

Pre-analysis Results

Before discussing the main findings, the judge fit statistics and information about the reliability of the CJ exercise are provided below.

Judge Statistics

The infit values (**Table 1**) were all within an acceptable range [0.5–1.5 as stated by Linacre (2002)]. The outfit values for judges 1, 2, 6, and 7 were below 0.5, suggesting that the observations were too predictable. As stated previously, this analysis was performed to examine judge behaviour; the analysis suggested that the judges were not misfitting the model to such an extent that they were affecting the estimates of script quality.

Comparative Judgment Script Measures

The Scale Separation Reliability was 0.8, indicating that the logit scale produced from the judgements could be considered reliable given the number of comparisons per script (30 comparisons per script for the 2018 scripts and 15 for the 2019 scripts). For high-stakes and summative assessments, a value of 0.9 is often considered desirable [cited in Verhavert et al. (2019)]. However, in a meta-analysis of CJ studies, Verhavert et al. (2019) found that this was achieved when there was a greater number of comparisons per script (26–37 comparisons).

The CJ measures are the logit values on this scale and indicate the relative overall judged performance of each script. When original candidate marks were compared to the CJ measures using Pearson's correlation, there was a strong relationship for the 2018 scripts [$r(38) = 0.92, p < 0.01$], indicating that candidate rank orders were similar for marking and the CJ judgements. The relationship is weaker for the 2019 scripts. The 2018 scripts were picked randomly, whereas the 2019 scripts were picked to exemplify certain characteristics and so could be considered "trickier" scripts to mark. This could explain the slightly weaker relationship between marks and measures and perhaps indicate that, for trickier scripts, there may be less similarity between marking and CJ. That the modified relationship [$r(38) = 0.83$, $p < 0.01$] was slightly weaker than the original [$r(38) = 0.86$, $p < 0.01$], which might indicate that the modifications are having an effect.

FINDINGS

We examined the CJ measures of the four features under consideration. The descriptive statistics are shown in **Table 2**, and the paired *t*-test results for each feature are shown in **Table 3**.

For each feature, the CJ measures of each variant were plotted against a script (**Figures 2–5**). Script numbers are listed on the x-axis; these range from 0 to 1, where 0 is the script with the lowest candidate mark and 9 is the script with the highest mark. As the scripts were chosen to be evenly spread across the mark scheme, we would expect the lines to go upward from left to right. They show whether any differences in measures between the two variants were consistent across the mark range.

Of the four features under consideration, the judges differed in whether they mentioned them during the observation (see **Table 4**). Since the observation was a "snapshot" of their judging, the presence or absence (rather than a count) of each feature was recorded. Only two judges (4 and 8) did not mention any of the four features during the observation. Handwriting, spelling, and missing responses were all reported in the survey responses. Appearance was not directly mentioned, but one participant mentioned "presentation." We will now examine each feature in turn.

Appearance

The descriptive statistics in **Table 3** show that, for the appearance feature, the mean CJ measures were quite similar for both the original ($M = 0.34$, $SD = 2.05$) and the modified ($M = 0.54$, $SD = 2.27$) variants. The range of measures was greater for the modified variant. The difference in mean measures was not significant [$t(9) = 0.29$, $p = 0.776$, $d = 0.09$], and in terms of approximate marks, the mean difference was less than one.

From **Figure 2**, we can see that the lines for the original and modified variants cross each other multiple times. This indicates that the modified CJ measure was higher for some of the scripts and the original variant was for others.

During the observation, half of the judges made reference to appearance features. The judges varied in how they expressed their comments, but they tended to be an observation or an aside that offered little explicit indication of whether this feature had influenced their judgements. Examples of appearance-specific comments included:

> *Few little crossings out, but things have been rewritten so that's OK (Judge 10).*
> *Again, lots of crossings out and rewriting things (Judge 10).*
> *Things at the side, little arrows on it (Judge 10).*
> *The crossing out in it doesn't help in terms of seeing a students work (Judge 5).*
> *You can see there straight away on the, [sic] on the first page we've got a crossing out (Judge 9).*

TABLE 1 | Judge fit: consistency with the Bradley–Terry model.

Judge	Number of judgements	Infit	Outfit
1	120	0.59	0.36
2	120	0.72	0.43
3	120	0.87	0.76
4	120	0.87	0.66
5	120	0.98	0.78
6	120	0.67	0.41
7	120	0.68	0.43
8	120	0.78	0.54
9	120	0.82	0.55
10	120	0.85	0.75

TABLE 2 | Descriptive statistics of the comparative judgment (CJ) measures for each of the four features.

Feature	Appearance		Handwriting		SPaG		Missing	
	Original	Modified	Original	Modified	Original	Modified	Original	Modified
N	10	10	10	10	10	10	10	10
Mean	0.34	0.54	−0.37	0.56	−0.29	−0.05	−0.01	−1.05
SD	2.05	2.27	2.44	2.17	1.70	1.85	1.90	1.48
Min	−2.64	−4.35	−4.64	−1.62	−3.03	−2.72	−3.24	−3.56
Max	4.92	4.27	2.87	4.75	2.81	3.76	2.87	0.76

TABLE 3 | Paired *t*-test results for the four features.

Feature	*N*	Mean difference (logits)	SE difference (logits)	*t*(9)	*p*	Cohen's d	Mean difference (approx. marks)
Appearance	10	0.19	0.65	0.29	0.776	0.09	0.95
Handwriting	10	0.93	0.42	2.21	0.054	0.70	4.65
SPaG	10	0.25	0.31	0.80	0.444	0.25	1.25
Missing	10	−1.04	0.39	−2.66	0.026	0.84	5.20

Taken together, the results suggest that the judges were not unduly influenced by script appearance; this is in contrast to the findings of Black et al. (2011) in a marking study. Some of the papers in this experiment could be considered very messy, so it is an encouraging sign that the judges were not influenced by this.

Handwriting

For handwriting, the mean difference in measures was just under one logit (0.93), indicating that, on average, scripts with improved handwriting [modified variant (M = 0.56, SD = 2.17)] had higher CJ measures than the original variant (M = −0.37, SD = 2.44). This result was borderline significant [$t(9)$ = 2.21, p = 0.054, d = 0.70]. The approximate effect of this was a mean difference in the marks of nearly five marks (4.65).

From **Figure 3**, we can see some defined trends. The modified scripts often had higher measures than the original scripts, particularly at the lower end of the script range. Where the modified scripts had lower measures, the difference was small.

Six of the judges made comments about handwriting; these were both positive and negative. The positive ones tended to describe being positively disposed to the writing, whereas the negative comments tended to be about not being able to read something and so not knowing if an answer was correct. Comments included:

> *Looking at this I can see it very clearly. The writing is lovely, which does help a marker (Judge 1).*
> *The handwriting is clear. It helps with, with, with, with [sic] the handwriting. Sometimes you can't, [sic] you can't decipher what they say, in which therefore ultimately hampers the marks (Judge 5).*
> *I can't read that. I think it might say hamstrings (Judge 9).*
> *I'm also looking at the actual writing itself. I have had a couple of them where the actual writing has been so bad, I couldn't actually read it despite going over it again and again (Judge 10).*
> *Just as an aside as well, not probably nothing to do with it. Sometimes you get put off by kids' writing. Or if it's really neat, yes, you, it can be quite positive towards especially on this when you are doing like reading through a paper. Whereas if I find if I'm doing all question ones [this refers to marking practice] and all questions twos you've not necessarily got that prejudice quite as obviously. Sometimes you gotta be careful with that, really, read kids marks (Judge 2).*

Interestingly, in the survey two judges acknowledged handwriting as a potential issue that they tried to ignore:

> *. . . I tried to ignore quality of handwriting.*
> *I did not focus on this when making my judgements, however, one area that could have had an impact was a students' handwriting. . ..*

The results suggest that the judges are influenced both positively and negatively by handwriting when making CJ judgements. Not being able to read an answer both increases the cognitive load and genuinely hampers the judges decision-making capability, so it is understandable if this causes problems. Being positively disposed toward a paper is of particular concern, as it is less tangible and so harder to correct. Recent research findings on *marking* have not found an effect of handwriting (Massey, 1983; Baird, 1998), so it was notable and concerning that a borderline effect was found in this context where it was hidden in a holistic judgement and, therefore, non-traceable.

Spelling, Punctuation, and Grammar

For SPaG, the mean CJ measures were quite similar for both the original (M = −0.29, SD = 1.70) and modified (M = −0.05, SD = 1.85) variants. The difference in mean measures was not significant [$t(9)$ = 0.80, p = 0.444, d = 0.25], and the mean difference in approximate marks was just over one.

Figure 4 shows that the measures were very close together for both features; only script 2 showed any sizeable difference (inspection of script 2 revealed no obvious reason for this). Again, the scripts varied as to whether the original or the modified variant had a higher measure. Four of the judges made comments on spelling, punctuation, or grammar and were not mentioned.

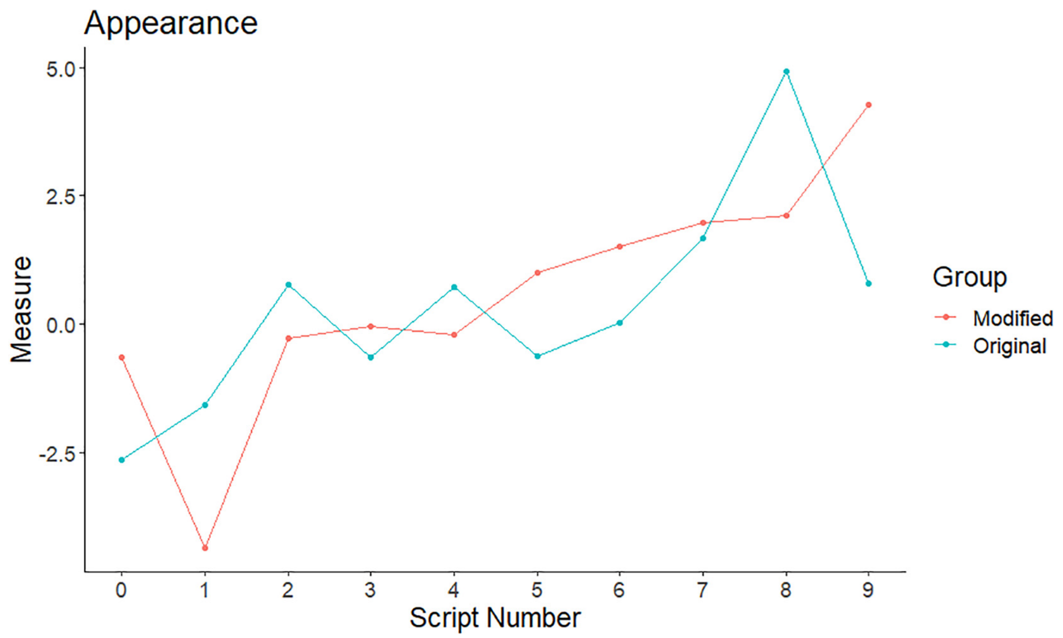


FIGURE 2 | Comparative judgment (CJ) measures for the original and modified script variants: appearance.

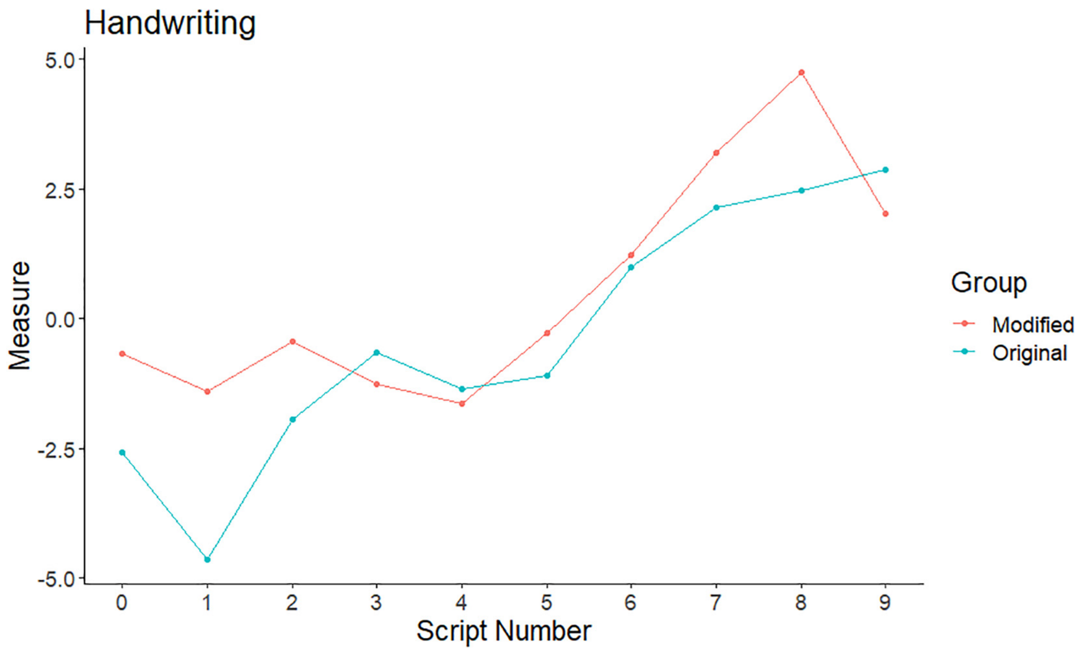


FIGURE 3 | CJ measures for the original and modified script variants: handwriting.

Comments include: "Phalanges, even though spelt wrong, is the correct answer" and "It's poor spelling."

SPaG appears to have very little influence on the judges' behaviour. This is encouraging, as it should not feature in the judgement of this paper. This is in line with other recent CJ studies (Bramley, 2009; Curcin et al., 2019) and, together, presents strong evidence that SPaG does not affect CJ judgements.

Missing

For missing, the mean difference in measures was just under one logit (−1.04), indicating that, on average, the original variant scripts (*M* = −0.01, *SD* = 1.90) had higher CJ measures than those where incorrect answers were replaced with a non-response (modified variant *M* = −1.05, *SD* = 1.48). The difference in mean CJ measures was significant

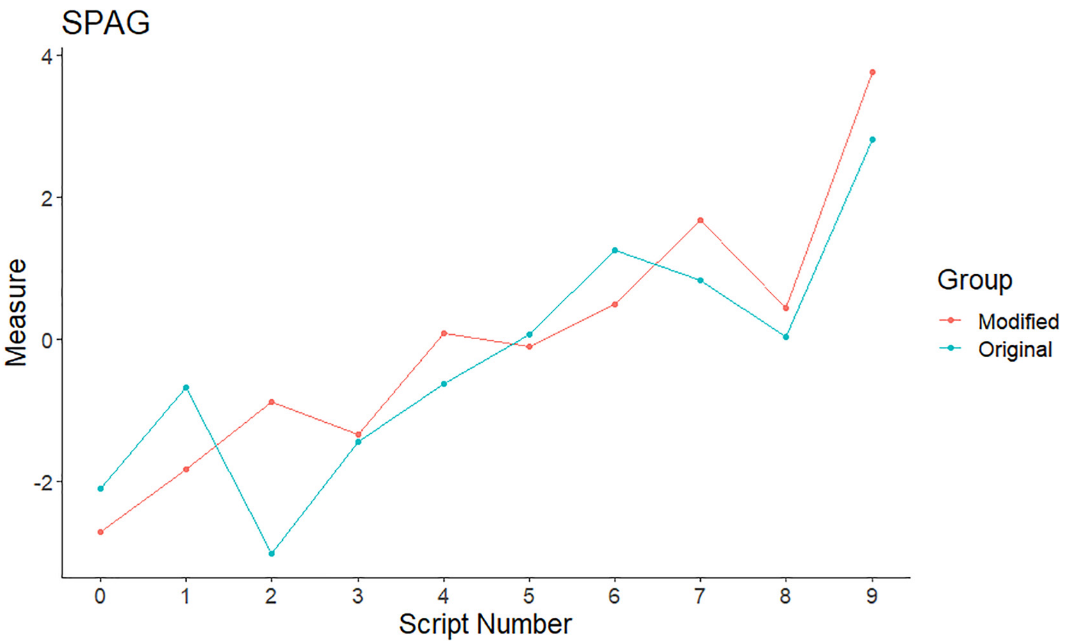


FIGURE 4 | CJ measures for the original and modified script variants: spelling, punctuation, and grammar (SpaG).

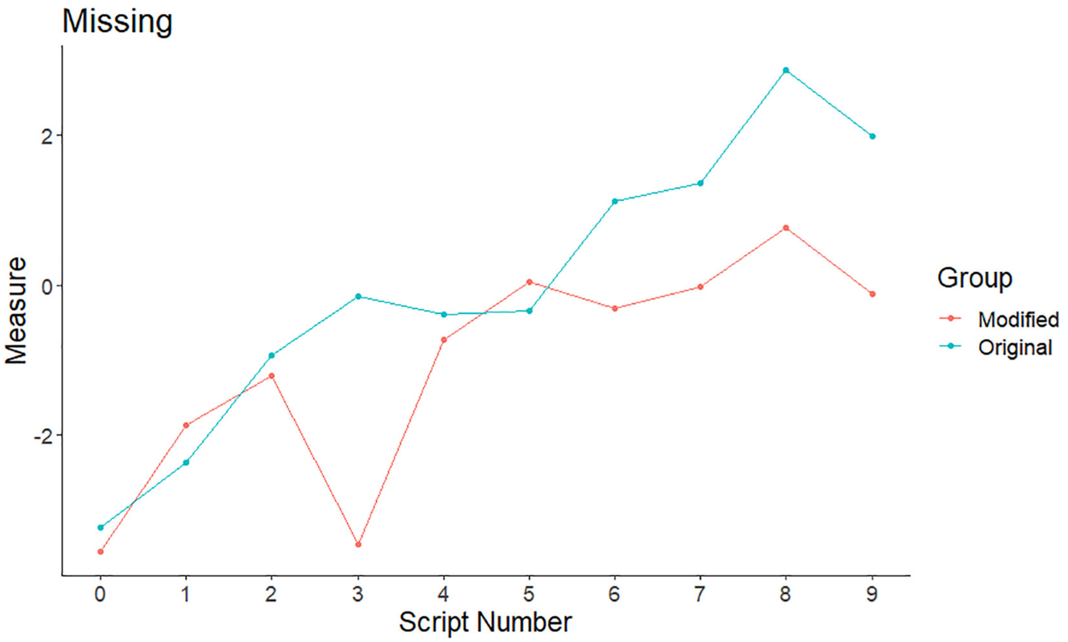


FIGURE 5 | CJ measures for the original and modified script variants: missing.

[$t(9) = -2.66$, $p = 0.026$, $d = 0.84$]. The approximate effect of this was a mean difference in marks of just over five marks (5.2).

Figure 5 shows that the differences in measures were quite pronounced, particularly at the higher mark range. Interestingly, a closer examination of the judgements on scripts 4 and 5 (where the measures are closer together) shows that the judges were split in their opinions, whereas for script 3 and scripts 6–9, the judges were in agreement.

Seven of the judges made comments about missing answers, some were an observation, some were about balancing the missing responses to the quality of other answers, and some were comments on several people leaving out a certain answer. Comments included:

> *Some missing responses, not many (Judge 9).*
> *I think this paper has a few more missing responses (Judge 9).*
> *Far too many questions being missed out, so looking like what has been done, been answered is not actually that bad (Judge 10).*
> *Just wanna check it against B though cos even though they have missed a lot of questions out what has been answered is pretty good (Judge 10).*
> *They filled everything in, but the quality isn't there (Judge 5).*
> *OK, even though there are mistakes, but there are some questions are missed. They are actually answering to more detail (Judge 6).*
> *There's gaps once again, is a big gap in knowledge there, which means, OK, we're going to be lowered down again in terms of ranking (Judge 5).*
> *Decided to leave that blank and they're not alone there (Judge 3).*

The challenge caused by balancing unattempted questions with the quality of the rest of the scripts and the further inspection required were also reported in the survey responses.

This evidence indicates that having missing answers, as opposed to incorrect answers, does influence the judges. In line with previous research (Bramley, 2009; Curcin et al., 2019), the missing responses appear to have a negative effect on CJ measures, suggesting that the judges were more negatively predisposed to a missing answer than an incorrect one. This is of concern, again because the holistic context makes it a hidden bias. It was encouraging, however, to see some of the judges acknowledging that, although there were missing answers, they should balance that with the content of what else was in the paper.

Other Potentially Construct-Irrelevant Features

When making their judgements, the judges mentioned a number of different features. The majority of these were ones we might expect and were relevant to the construct or marking practice, e.g., whether the question was answered, the use of terminology or keywords, the use of supporting examples, giving the benefit of doubt, the vagueness of answers, and a candidate's level. The survey responses corroborated this. Construct-relevant strategies cited in the survey were "number of correct answers," "knowledge," "the level of detail," "use of technical language," and the use of "practical examples."

However, two features were mentioned, both in the think-aloud procedure and the survey responses, that could potentially be considered as construct-irrelevant; the use of exam technique and whether the candidates wrote in sentences. Both features were considered positive. Judges 4, 9, and 10 referred to "examination technique" which included things like underlining or ticking keywords in the question and writing down acronyms, e.g., "So we've got a bit of a plan up here with [. . .] circling and underlining key points, which is what I like. This candidate's obviously thinking about their response." Only one judge (10) made reference to candidates writing in sentences and did this multiple times e.g., "They have tried to write in sentences, which is good." It is encouraging that no other potentially construct-irrelevant features were mentioned in the observations or surveys, which hopefully implies that they were not being attended to.

CONCLUSION AND RECOMMENDATIONS

This study sought to explore judges' decision-making in CJ, specifically to focus on one aspect of the validity of these judgements: whether judges were attending to construct-irrelevant features. As noted earlier, the validity of CJ is comprised of both the holistic nature of decision-making and a shared consensus of judges. Focus on construct-irrelevant features could impact both of these elements.

The study was conducted within an awarding organisation; the particular context was set within a series of studies trialling a new method of maintaining examination standards involving CJ. Judgements in this context are cognitively demanding, and there is a possibility that judges may attend to superficial features of the responses they are comparing. Our research question was as follows: Are judges influenced by the following construct-irrelevant features when making CJ decisions in a standard maintaining context?

- Appearance.
- Handwriting.
- SPaG.
- Missing response vs. incorrect answer.

We investigated this using a mixed-method design, triangulating the results from a quantitative element formed from an empirical experiment and a qualitative element formed from judge observations and survey responses. We found that the different sources of evidence collected in the study supported each other and painted a consistent picture.

The appearance and SPaG features did not appear to affect judges' decision-making. For SPaG, this is in line with other recent CJ studies (Bramley, 2009; Curcin et al., 2019) despite some of the judges mentioning spelling in their observation/survey responses. For appearance, this had not been investigated in a CJ context; however, in a marking study, Black et al. (2011) found that appearance features did interfere with marking strategies. This study suggested that this interference

TABLE 4 | Each judge's mentions of the four features during observation.

Judge	Appearance	SPaG	Handwriting	Missing
1			✓	✓
2	✓	✓	✓	
3	✓		✓	✓
4				
5	✓		✓	✓
6				✓
7		✓		
8				
9	✓	✓	✓	✓
10	✓	✓	✓	✓

was not strong enough to affect judging outcomes in this context. For both features, this is a positive outcome particularly given that the scripts were either very untidy or had many SPaG errors.

However, handwriting (to some extent) and missing responses vs. incorrect answers did appear to affect judges' decision-making. For handwriting, this was particularly for scripts at the lower end of the mark range. Recent research findings on *marking* have not found an effect of handwriting, so it is notable and concerning that a borderline effect was found in this context where it was hidden in a holistic judgement and, therefore, non-traceable. However, the scripts used were difficult to read, so the influence of this feature may be restricted to these extreme cases.

For missing responses, the effect was found at the mid- to high-end of the mark range. This is in line with previous research (Bramley, 2009; Curcin et al., 2019). This is of concern, again, because the holistic context makes it a hidden bias. Why judges should be more negatively influenced by a missing response than an incorrect answer is an interesting question. When viewing a script, the presence of missing responses is immediately apparent to a judge and, thus, could be treated as a quick differentiator of quality. A number of "gaps" in a script may suggest gaps in a student's knowledge, perhaps more so than a number of incorrect answers. There may be some influence of an incorrect answer suggesting that the students had tried to answer. These explanations are speculative and would need investigation.

Both handwriting and missing responses were directly mentioned by the judges, and some of the judges offered strategies to reduce any influence. In the case of handwriting, the strategy was to try and ignore it; in the case of missing responses, it was to attempt to balance the missing responses and content. In CJ, unlike marking, there is no audit trail of decision-making, so the influence of these features is not apparent and cannot be corrected after the event. Thus, judges attending to superficial or construct-irrelevant features are a threat to the validity of the CJ standard maintaining process and could compromise outcomes. In this context, however, there is scope to mitigate any effect.

Practical solutions for standard maintaining would be to (i) avoid using scripts with lots of missing responses or with hard-to-read handwriting or (ii) confirm that the scripts selected are representative of all scripts on the same mark in these respects. Scripts with many missing responses could be identified programmatically; however, handwriting would require visual inspection.

The observation and survey data indicated that the majority of the features attended to were based on construct-relevant features, e.g., whether a question was answered, the use of terminology or keywords, the use of supporting examples, etc. We found that the judges generally did not attend to other construct-irrelevant features, which is reassuring. Two other features were mentioned: (i) the examination technique, the presence of which was seen as a positive, and (ii) writing in sentences, which was noted as something to look for. The use of either was not widespread.

As noted earlier, in holistic decision-making, judges decide what constitutes good quality, and this conceptualisation determines their rank order of the scripts. The "use of CJ is built on the claim that rooting the final rank order in the shared consensus across judges adds to its validity" (van Daal et al., 2019, p. 3). This shared consensus is more than agreement; it is about a "shared conceptualisation" of what they are judging as the "judges' collective expertise defines the final rank order" (p.3). The judge statistics indicated that the judges did achieve consensus, but was it an appropriate consensus? Consensus is good if an appropriate range of aspects are considered, but less so if judges focus on a narrow range of, or incorrect, features.

In our experiment, we observed the judges, so we know what features and strategies they reportedly attended to, which, typically, we would not do. While we anticipate that the judges will use their knowledge of the assessment and their marking experience to make their judgements (Whitehouse, 2012), the use of CJ does place a lot of faith in the judges judging how we expect or would like them to. Without adequate training, we cannot make this assumption, particularly in the standard maintaining scenario where we are expecting the judges to set aside their many years of marking practice and potentially apply a new technique.

It is recommended that judges have training on making CJ decisions that involves practice, feedback, and discussion. Training on awareness of construct-irrelevant features could be introduced. However, it would need to be implemented cautiously and tested to ensure that judges do not overcompensate and cause problems in the other direction. In terms of future research activities, it is recommended that researchers meet with judges before a study to explain the rationale and ensure that judges know what is expected of them. Practice activities would also be useful. For CJ more generally, while previous research (e.g., Heldsinger and Humphry, 2010; Tarricone and Newhouse, 2016) has cited the small amount of training needed as one of the advantages of CJ; this study shows that it might be time to revisit the topic.

As training was a key recommendation, it would be valuable to both replicate this study and explore these findings further in studies where training on how to make holistic decisions was given to judges. As appearance and SPaG seemed not to influence judges, research attention could be directed to handwriting and missing responses, and this would give more freedom in qualification selection as papers that directly assessed SPaG could be included.

While this study was set in a specific standard maintaining context involving cognitively demanding judgements, there are applications to a wider CJ context. Particularly, the lack of influence of SPaG and appearance features – that these were shown not to have an effect in this highly demanding context should be reassuring for other assessment contexts in which these features do not form part of the assessment construct. For handwriting and missing responses, where the option exists to include/exclude scripts, hard-to-read scripts or scripts with missing responses, could be avoided.

Before concluding, it is important to note some limitations of the study. First, the number of scripts in each feature category was quite small at only 10, meaning the power to detect a difference between the variants was quite low. However, despite this, differences were detected. Second, the scripts were selected or constructed to exemplify particular features so they could be considered to be less typical or perhaps "problematic." Thus, these results hold for stronger instances of the features in question, and a less striking instance of the feature may have less or no effect.

This study contributes to existing research, both on which aspects guide judges' decisions when using CJ and on the impact of judges attending to construct-irrelevant features. In summary, the study did reveal some concerns regarding the validity of using CJ as a method of standard maintaining. This was with respect to judges focusing on superficial, construct-irrelevant features, namely, handwriting and missing responses. These findings are not necessarily a threat to the use of CJ in standard maintaining, as with careful consideration of the scripts and appropriate training, these can potentially be overcome.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because open access consent was not obtained. Requests to access the datasets should be directed to LC, Lucy.chambers@cambridge.org.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

LC and EC contributed to conception, design, and set up of the study. LC performed the statistical and qualitative analysis and wrote the manuscript. Both authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We would like to acknowledge the 40 volunteers who rewrote the candidate scripts.

REFERENCES

Baird, J. A. (1998). What's in a name? Experiments with blind marking in a-level examinations. *Educ. Res.* 40, 191–202. doi: 10.1080/0013188980400207

Benton, T., Cunningham, E., Hughes, S., and Leech, T. (2020a). *Comparing the Simplified Pairs Method of Standard Maintaining to Statistical Equating Cambridge Assessment Research Report*. Cambridge: Cambridge Assessment.

Benton, T., Leech, T., and Hughes, S. (2020b). *Does Comparative Judgement of Scripts Provide an Effective Means of Maintaining Standards in Mathematics? Cambridge Assessment Research Report*. Cambridge: Cambridge University Press.

Benton, T., Gill, T., Hughes, S., and Leech, T. (2022). *A Summary of OCR's Pilots of the use of Comparative Judgement in Setting Grade Boundaries. Research Matters: A Cambridge University Press & Assessment Publication, 33.* Cambridge: Cambridge University Press. 10–30.

Black, B., Suto, I., and Bramley, T. (2011). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement. *Assess. Educ.*18, 295–318. doi: 10.1080/0969594X.2011.555328

Bradley, R. A., and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39, 324–345. doi: 10.2307/2334029

Bramley, T. (2007). "Paired comparison methods," in *Techniques for Monitoring the Comparability of Examination Standards*, eds P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (London: QCA), 246–300.

Bramley, T. (2009). "The effect of manipulating features of examinees' scripts on their perceived quality," in *Proceedings of the Annual Conference of the Association for Educational Assessment–Europe (AEA-Europe)*, Malta.

Chase, C. I. (1983). Essay test scores and reading difficulty. *J. Educ. Meas.* 20, 293–297. doi: 10.1111/j.1745-3984.1983.tb00207.x

Craig, D. A. (2001). *Handwriting Legibility and Word-Processing in Assessing Rater Reliability*. Master's thesis. Champaign, IL: University of Illinois.

Crisp, V. (2013). Criteria, comparison and past experiences: how do teachers make judgements when marking coursework? *Assess. Educ.* 20, 127–144. doi: 10.1080/0969594X.2012.741059

Curcin, M., Howard, E., Sully, K., and Black, B. (2019). *Improving Awarding: 2018/2019 Pilots*. Coventry: Ofqual.

Davies, B., Alcock, L., and Jones, I. (2021). What do mathematicians mean by proof? A comparative-judgement study of students' and mathematicians' views. *J. Math. Behav.* 61:100824. doi: 10.1016/j.jmathb.2020.100824

Heldsinger, S., and Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *Aust. Educ. Res.* 37, 1–19. doi: 10.1007/BF03216919

Lesterhuis, M., van Daal, T., Van Gasse, R., Coertjens, L., Donche, V., and De Maeyer, S. (2018). When teachers compare argumentative texts. Decisions informed by multiple complex aspects of text quality. *L1 Educ. Stud. Lang. Lit.* 18, 1–22. doi: 10.17239/L1ESLL-2018.18.01.02

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Meas. Trans.* 16:878.

Massey, A. (1983). The effects of handwriting and other incidental variables on GCE 'A' level marks in English literature. *Educ. Rev.* 35, 45–50. doi: 10.1080/0013191830350105

Meadows, M., and Billington, L. (2005). *A Review of the Literature on Marking Reliability*. London: National Assessment Agency.

Messick, S. (1989). Meaning and values in test validation: the science and ethics of assessment. *Educ. Res.* 18, 5–11. doi: 10.3102/0013189X018002005

Plackett, R. L. (1975). The analysis of permutations. *J. R. Stat. Soc. Ser. C Appl. Stat.* 24, 193–202. doi: 10.2307/2346567

Pollitt, A. (2012a). Comparative judgement for assessment. *Int. J. Technol. Des. Educ.* 22, 157–170.

Pollitt, A. (2012b). The method of adaptive comparative judgement. *Assess. Educ.* 19, 281–300. doi: 10.1080/0969594x.2012.665354

Pollitt, A., and Crisp, V. (2004). "Could comparative judgements of script quality replace traditional marking and improve the validity of exam questions?," in *Proceedings of the British Educational Research Association Annual Conference*, Manchester.

Shaw, S., Crisp, V., and Johnson, N. (2012). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assess. Educ.* 19, 159–176. doi: 10.1080/0969594X.2011.563356

Stewart, M. F., and Grobe, C. H. (1979). Syntactic maturity, mechanics of writing, and teachers' quality ratings. *Res. Teac. English* 13, 207–215.

Suto, I., and Novaković, N. (2012). An exploration of the examination script features that most influence expert judgements in three methods of evaluating script quality. *Assess. Educ.* 19, 301–320. doi: 10.1080/0969594X.2011.592971

Tarricone, P., and Newhouse, C. P. (2016). Using comparative judgement and online technologies in the assessment and measurement of creative performance and capability. *Int. J. Educ. Technol. High. Educ.* 13:16. doi: 10.1186/s41239-016-0018-x

Turner, H. L., van Etten, J., Firth, D., and Kosmidis, I. (2020). Modelling rankings in R: the plackett-luce package. *Comput. Stat.* 35, 1027–1057. doi: 10.1007/s00180-020-00959-3

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., and De Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assess. Educ.* 26, 59–74. doi: 10.1080/0969594X.2016.1253542

Verhavert, S., Bouwer, R., Donche, V., and De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assess. Educ.* 26, 541–562. doi: 10.1080/0969594X.2019.1602027

Whitehouse, C. (2012). *Testing the Validity of Judgements about Geography Essays Using the Adaptive Comparative Judgement Method*. Manchester: AQA Centre for Education Research and Policy.

Wright, B. D., and Masters, G. N. (1990). Computation of OUTFIT and INFIT statistics. *Rasch Meas. Trans.* 3, 84–85.

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 28 April 2022
doi: 10.3389/feduc.2022.803040

The Accuracy and Validity of the Simplified Pairs Method of Comparative Judgement in Highly Structured Papers

Tony Leech†, Tim Gill*†, Sarah Hughes and Tom Benton*

Assessment Research Division, Cambridge University Press & Assessment, Cambridge, United Kingdom

Comparative judgement (CJ) is often said to be more suitable for judging exam questions inviting extended responses, as it is easier for judges to make holistic judgements on a small number of large, extended tasks than a large number of smaller tasks. On the other hand, there is evidence it may also be appropriate for judging responses to papers made up of many smaller structured tasks. We report on two CJ exercises on mathematics and science exam papers, which are constructed mainly of highly structured items. This is to explore whether judgements processed by the simplified pairs version of CJ can approximate the empirical difference in difficulty of pairs of papers. This can then be used to maintain standards between exam papers. This use of CJ, not its other use as an alternative to marking, is the focus of this paper. Within the exercises discussed, panels of experienced judges looked at pairs of scripts, from different sessions of the same test, and their judgements were processed *via* the simplified pairs CJ method. This produces a single figure for the estimated difference in difficulty between versions. We compared this figure to the difference obtained from traditional equating, used as a benchmark. In the mathematics study the difference derived from judgement *via* simplified pairs closely approximated the empirical equating difference. However, in science, the CJ outcome did not closely align with the empirical difference in difficulty. Reasons for the discrepancy may include the differences in the content of the exams or the specific judges. However, clearly, comparative judgement need not lead to an accurate impression of the relative difficulty of different exams. We discuss self-reported judge views on how they judged, including what questions they focused on, and the implications of these for the validity of CJ. Processes used when judging papers made up of highly structured tasks were varied, but judges were generally consistent enough. Some potential challenges to the validity of comparative judgement are present with judges sometimes using re-marking strategies, and sometimes focusing attention on subsets of the paper, and we explore these. A greater understanding of what judges are doing when they judge comparatively brings to the fore questions of judgement validity that remain implicit in marking and non-comparative judgement contexts.

Keywords: comparative judgement, pairwise comparisons, standard maintaining, structured exams, educational assessment, simplified pairs

OPEN ACCESS

Edited by:
Sven De Maeyer,
University of Antwerp, Belgium

Reviewed by:
Liesje Coertjens,
Catholic University of Louvain, Belgium
Fien De Smedt,
Ghent University, Belgium

Correspondence:
Tony Leech
anthony.leech@cambridge.org
Tim Gill
tim.gill@cambridge.org

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement,
a section of the journal Frontiers in Education

Received: *27 October 2021*
Accepted: *06 April 2022*
Published: *28 April 2022*

Citation:
Leech T, Gill T, Hughes S and Benton T (2022) The Accuracy and Validity of the Simplified Pairs Method of Comparative Judgement in Highly Structured Papers. Front. Educ. 7:803040. doi: 10.3389/feduc.2022.803040

INTRODUCTION AND CONTEXT

High stakes exams in England have since 2011 used a standard maintaining approach called 'comparable outcomes', the aim of which is to avoid grade inflation by using statistical approaches to set grade boundaries to ensure roughly the same percentage of students get the same grade each year.[1] Judgement of performance through inspection of scripts is limited to a small sample of scripts, typically between six and ten, on marks near key grade boundaries. The comparable outcomes method has been criticised for its inability to reflect any genuine change in performance year on year, while the script inspection element, due to how limited it is, has been criticised for a lack of rigour and a lack of independence from the statistical approaches. In response to these criticisms, Ofqual, the Regulator of educational qualifications for England, in 2019 invited the exam boards in the United Kingdom to discuss and investigate the possibility of using comparative judgement (CJ) evidence in maintaining exam standards (Curcin et al., 2019, p. 14) as this could allow better use of evidence based on judgements of candidate scripts.

Using CJ to maintain exam standards typically involves exam scripts from the current exam for which standards are being set and scripts from a previous exam representing the benchmark standard that is to be carried forward. Judges are presented with pairs of student exam scripts – typically one from each exam – and make decisions about which is better. Many judgements are made by many judges. A statistical model such as the Bradley-Terry model (Bradley and Terry, 1952, p. 325) is then used to convert these judgements into a measure of script quality. The measures for each script from both exams are located on the same scale allowing the mapping of scores (and boundary marks for different grades) from one exam to the other, thus allowing a boundary mark on the benchmark exam to be equated to the new exam (Bramley, 2005, p. 202). A simplified version of this method has also been developed at Cambridge University Press & Assessment, as described in Benton et al. (2020, p. 5). In this method, called simplified pairs, the mapping of scores between different tests is undertaken without the need to estimate values on a common scale by fitting a statistical model. This makes it more efficient, as scripts only need to appear in one comparison, rather than many. Note that this method can only be used as a means to find a mapping between two existing mark scales. Unlike other CJ approaches, it does not provide a fresh ranking of the exam scripts included in the study (see, Benton, 2021, for further details).

In 2019 Oxford, Cambridge and RSA Examination (OCR) – one of the exam boards which deliver high stakes exams in England, and part of Cambridge University Press & Assessment – correspondingly launched a programme of research which aimed to evaluate the effectiveness of using comparative judgement to maintain standards in exams. The programme eventually comprised 20 CJ exercises across several subjects and qualification types and overall outcomes are recorded in Benton et al. (2022). The present article, in focusing specifically on the two exercises on highly structured papers, and exploring insights in more detail, makes a contribution distinct from that work.

Comparative judgement requires that judges make *holistic* judgements of student work. Much previous research on the use of CJ in awarding has focused on examinations requiring essay-type responses (e.g., Gill et al., 2007, p. 5; Curcin et al., 2019, p. 10). CJ has been successfully used to "scale performances by students in creative writing essays, visual arts, philosophy, accounting and finance, and chemistry (laboratory reports)", according to Humphry and McGrane (2015, p. 459) who assert that it is promising for maintaining standards on assessments made up of extended tasks. For this reason, initially the OCR research programme trialled assessments made of extended tasks where holistic judgements might be seen to be more appropriate, e.g., Sociology, English Language and English Literature.

However, OCR were interested in the possibility of applying the same CJ standard-setting methods across all subjects, including science, technology, engineering, and mathematics (STEM) subjects which tend not to include extended tasks. CJ has not been studied as thoroughly in relation to STEM subjects, or subjects utilising extended tasks. Consequently, exercises in both mathematics and science were set up, and are reported on here. The present article aims to answer questions about the nature of holistic judgements and whether judges can make them on highly structured papers. This will help to inform debate on the future use of CJ for maintaining standards in STEM subjects.

The insights from this paper may also be valuable for those interested in maintaining exam standards in other, non-United Kingdom, education systems that utilise high-stakes external tests where it is important to maintain standards. The procedures discussed in the present article were developed in and for the United Kingdom context, in which there is a need to equate standards of assessments from year to year. Crucially, in this context, more established statistical equating methods, such as pre-testing of items, are not available, as items are created anew for each year and are not released in advance of the exams being sat in order to preserve the confidentiality of the assessment. Note that in what follows, it is uses of CJ for standard maintaining that are discussed; the use of CJ as an alternative to marking is not a focus here. In the discussed exercises, all the scripts used had been marked – the goal is the maintenance of standards between assessments set in different years.

The paper starts with a brief literature review and a discussion of prior findings from OCR trials for non-STEM subjects (which utilised mostly more extended tasks) including surveys of judges taking part in these trials. Following that, we present the findings of two CJ exercises in STEM subjects (mathematics and science), as well as the outcomes of surveys of their judges. Our discussion and conclusions focus on the accuracy and validity of CJ for judging highly structured exam papers such as those used to assess STEM subjects in United Kingdom high stakes exams. We

[1] Ofqual (2017) described, "if the national cohort for a subject is similar (in terms of past performance) to last year, then results should also be similar at a national level in that subject".

then consider implications for decision-making about whether CJ is a suitable way to maintain exam standards in subjects using highly structured papers.

Literature Review

Comparative Judgement and Judgemental Processes

Literature identifies features which may impact the accuracy and validity of judgements and judges' ability to make judgements. These can relate to the processes judges use, the questions (or parts of scripts) that they attend to, and whether they are able to make holistic judgements or conversely just end up re-marking the papers and adding up the marks. This section will explore what we know currently about judgemental processes in assessment.

Work on the cognitive processes used by judges of exam scripts has been pioneered by Cambridge Assessment researchers, with a substantial series of linked research projects in the 2005–2010 period central to that (e.g., Suto and Greatorex, 2008, p. 214; Suto et al., 2008, pp. 7–8; Crisp, 2010a, p. 3). The research area is a subset of the field of judgement and decision-making, in which there has been psychological research under various paradigms. Areas such as "what information people pay attention to", the heuristics and biases they face, and the role of the behavioural and social, were explored, as were assessments of the sequences of mental processes undertaken when making decisions. Much of this research, however, focused on marking. Through think-aloud sessions, observation and interview, the processes used in marking, such as scrutinising, elaborating and scanning, were described (e.g., Suto et al., 2008, p. 7; Crisp, 2010b). Crisp (2008) found in a marking study that most aspects of the candidate work noted by examiners related to relevant content knowledge, understanding and skills. As discussed, the present article considers the case for CJ for standard maintaining purposes, not as an alternative to marking.

Where there are several questions that must be considered in a script it is possible that judges may only pay attention to a subset (Verhavert et al., 2019). For Verhavert et al. (2019) the structure of a task impacts on both the reliability and the complexity of a CJ exercise for judges. Similarly, in a study of different CJ approaches to making grading decisions in a biology exam, Greatorex et al. (2008, pp. 4–5) report that it was clear that not all questions received equal consideration. The researchers found from analysis of which questions judges referenced as those that they focused on the most, that the same question (a long-answer question with more marks than any others on the paper) was referenced for all methods. Crucially, however, this long answer question empirically discriminated poorly, suggesting that judges are not good at determining which questions they should be focusing on due to their greater discriminatory power. They concluded that what judges across these methods focus on were "some key questions but not necessarily the most useful ones" (p. 9). Greatorex (2007, p. 9), in reviewing wider literature, highlights that "experts are good at knowing what they are looking for but they are not good at mentally combining information".

Using CJ for maintaining standards between tests will require that judges compare performances on different tests including different questions, e.g., a response to the 2018 exam and the 2019 exam. These two exams will intend to assess the same constructs to the same standard, but the difficulty of the particular questions and therefore exams varies between years. (In the United Kingdom's exam systems papers are not pre-tested so the difficulty of the items is not known before papers are sat). Judgements between different exam papers require that judges can take some account of these differences in their decision-making. Black (2008, p. 16) found that judges in a CJ exercise tended to suggest that comparing scripts where the candidates were answering different questions – "because the papers under comparison were different in different years" – was "fairly difficult". Judges noted that they frequently had to remind themselves what the candidates were writing about, and that it is difficult to make like-with-like comparisons in this context.

Baird (2007, p. 142) raised the concern that "examiners cannot adequately compensate in their judgements of candidates' work for the demands of the question papers". The concern is that, as suggested by Good and Cresswell (1988, p. 278), subject experts will be more impressed by a candidate achieving a high score on an easy paper than by a candidate achieving a (statistically equivalent) lower score achieved on a harder paper. An experiment presented by Benton et al. (2020, p. 21) for an English literature examination appears to suggest that this concern is not always justified, as in that case the CJ method meant that judges were able to appropriately make allowances for paper difficulty. This paper extends this work to mathematics and science exams when grade boundaries are set using CJ.

Humphry and McGrane (2015, p. 452) highlight that judging between responses to different exam questions, potentially of different difficulty, across several assessment criteria, can increase cognitive load to the extent that the task becomes difficult and potentially unreliable. This therefore brings out the question of potential re-marking of each individual question – "rather than making a holistic judgement" – and then just adding up the scores (though this means the difficulty of each paper is not accounted for). Leech and Chambers (2022) found that when judging a physical education exam, judges varied in their approach. Some re-marked the scripts, only one (of six) marked purely holistically, and the others combined both approaches. The level of re-marking of each question observed suggests that judgements were only partially, if at all, holistic.

Another issue in relation to what judges attend to is the question of construct-irrelevant features. Bramley (2012, p. 18) carried out an experiment into whether manipulating features of scripts that did not alter the marks, such as quality of written response and proportion of missing to incorrect responses, changed judges' views of script quality. The two largest effects were seen by changing the proportion of marks gained on items defined as testing "good chemistry" knowledge, (Bramley, 2012, p. 19) where scripts with a higher proportion appeared better on average, and replacing incorrect with missing responses, where scripts with missing responses appeared worse on average. The implication is that the decision on relative quality is affected by the makeup of the scripts chosen. More recent work on this (Chambers and Cunningham, 2022) found that replacing incorrect answers with missing answers affected judges' decision-making. Scripts with missing responses, rather than zeroes, received statistically significantly lower script measures on average. If judges are looking at construct-irrelevant features, this is a threat to the validity of the CJ awarding process.[2] Chambers and Cunningham found that other construct-irrelevant features of spelling, punctuation, grammar, and appearance (e.g., crossings-out and text insertions and writing outside of the designated answer area) did not impact judges' decisions, however.

Comparative Judgement Specifically in STEM Subjects

STEM subjects have been previously investigated as part of CJ exercises. For example, the accuracy of holistic judgements in history (non-STEM) and physics (STEM) was investigated by Gill and Bramley (2013, p. 310). In this study, examiners made three different kinds of judgements. These were: absolute judgements (that is, was the script worthy of the grade or not?), comparative judgements (of which script is better), and judgements of their own confidence in their other judgements. In both subjects, relative judgements were more accurate than absolute ones, and judgements the examiners were 'very confident' in were more accurate than other judgements. However, in physics, the further apart two scripts were in terms of overall mark the greater the likelihood of a correct relative judgement, but in history the link was weaker. This may suggest that in STEM there are more "right answers" and less scope for legitimate differences in judge professional judgement.

Jones et al. (2015, p. 172) used CJ to successfully assess mathematical problem solving. They highlighted that CJ was more useful when judging mathematics if longer, more open-ended tasks were used. In a similar manner, Humphry and McGrane (2015, p. 457) described paired CJ comparisons as "likely to be more suitable for extended tasks because they allow students to show a range of abilities in a single and coherent performance, which can be compared holistically". However, in examinations assessing STEM subjects – at least as currently designed in the United Kingdom – there are typically not a small number of extended tasks, but many shorter answer questions. As Jones et al. (2015) indicate, this is not an intrinsic feature of STEM assessments, but is generally the case, at least in the United Kingdom. STEM assessments and highly structured exams are not synonymous. This means that it is not whether a subject is STEM or not that determines whether it is appropriate for use in CJ, but how structured the exam is. In other words, item design, not item content, is the issue at stake. This paper will therefore be discussing the issue in this way.

Findings From OCR Trials of Assessments Based on Extended Tasks

Initially the OCR programme looked at assessments where holistic judgements might be more straightforward, as tasks are generally extended response and fewer in number. The programme investigated, among others, Sociology, English Language and English Literature. The precision of the outcomes of these exercises was high, with standard errors (which indicate the precision of the grade boundary estimates) of between 1.5 and 2.5 on each test – i.e., typically a confidence interval of ±1.2 marks on the test (Benton et al., 2022).

Point biserial correlations[3] demonstrate the association between the CJ judgement and the original marks given to each script. For exams comprising extended responses these were between 0.34 and 0.52 - encouraging figures. Further trials included exam papers with a mix of more and less structured tasks e.g., Geography, Business Studies, Enterprise and Marketing, Child Development, and Information Technologies (Benton et al., 2022). Outcomes of these CJ exercises were as accurate as with all the exercises using extended response question papers, with standard errors between 1.4 and 2.7. Consequently, we thought perhaps CJ exercises on papers made up mainly of highly structured tasks would be equally reliable.

The judges of these exercises were also asked about how they make their decisions. OCR judges differed in their views of whether it was at all straightforward to compare responses to tests from different series (i.e., those with different sets of questions, albeit likely similar in form). While many judges felt that they were able to do this, another was "not sure it was possible" and some papers were described as "apples and pears". This corresponds to the insights of Black (2008, p. 16), mentioned earlier.

A further question of interest is that of how judges decide between scripts which each demonstrate different legitimate strengths to different degrees. Many judges in these trials suggested they had difficulty deciding between, for instance, scripts with greater technical accuracy and greater "flair", or scripts with strengths in reading and strengths in writing, and so on. This was clearly a challenge for many judges (roughly a third in these trials).

A notable number of judges responded that they were making judgements primarily on certain long-answer questions. The validity of this approach can be challenged. On the one hand, candidates should answer the whole script, and awards are based on all responses. On the other, these questions are worth more marks and are likely therefore to contribute more to rank order determinations on marks as well as by judgement. They might be seen to demonstrate more true ability. However, most judges who said that they judged mainly on long answer questions said they did so *as these questions were worth more marks*, not because they were seen as intrinsically stronger determiners of quality.

Judges in these trials had not necessarily internalised an idea of "better" that is distinct from *what the mark scheme says should be credited.* What they were effectively asking for was some measure of standardisation. Even those judges who said that the constructs they were judging resided in their minds, not in the mark scheme, suggested that this was because of their experience of marking. This then calls into question the idea that judges can make holistic

[2] This threat is partly mitigated by the fact that, in the CJ experiments discussed here, scripts with more than 20% of their responses missing were excluded.

[3] This is the Pearson correlation between judges' decisions (expressed as values of 0 or 1) and the mark differences between the scripts being compared.

judgements separate from marks, or at least that they can be confident in what they are doing. On the other hand, other judges suggested that CJ made them more thoughtful and deep judgements of quality.

RESEARCH QUESTIONS

In order to assess the suitability of the simplified pairs method of comparative judgement for accurately estimating the true difference in difficulty between pairs of highly structured papers, and to investigate the nature of decisions made by judges, we defined two research questions. These were:

> RQ1 (the accuracy question): "Can comparative judgement estimate the true difference in difficulty between two exam papers comprising many highly structured tasks?"
>
> RQ2: (the validity question): "How do judges make comparative judgements of students work from exam papers comprising many highly structured tasks, and what validity implications does understanding their processes have?"

In the main, RQ1 was addressed using the results of the CJ exercises, while RQ2 is addressed *via* insights from follow-up surveys of the judges involved in the exercises.

METHOD

Comparative Judgement Exercises

The first aim of both studies reported in this paper was to assess whether the simplified pairs method of comparative judgement could accurately estimate the true difference in difficulty between two exam papers (as determined by statistical equating). If they can, this means there is the potential for the method to be used in standard maintaining exercises, where the difference in difficulty between last year's paper and this year's is fundamental.

In the simplified pairs method (see Benton et al., 2020, p. 5), judges undertake many paired comparisons and decide which of each pair is better, in terms of overall quality of work. For example, there might be six judges who each make 50 comparisons between pairs of scripts (one from each exam paper), with the difference in marks (from the original marking) for each pair varying between 0 and 25 marks. In about half the comparisons, the paper 1 script will have the higher mark and about half the time the paper 2 script will have the higher mark. Each script should only appear in one paired comparison, so 300 different scripts from each paper will be required.

For each paired comparison, the number of marks given to each script is recorded, as well as which script won the comparison. This is so that we can determine the relationship between the mark difference and the probability that script A (from paper 1) beats script B (from paper 2). This relationship is then used to answer the following question:

Suppose a script on paper 1 has been awarded a score of x. How many marks would a script from paper 2 need to have a 50% chance of being judged superior?

TABLE 1 | Relationship between mark difference and probability of superiority.

Mark difference (paper 2 – paper 1)	No. of paired comparisons	% where paper 2 judged superior
−1	10	25
0	8	50
1	9	55
2	10	40
3	7	71
etc.		

If we have many paired comparisons for each mark difference, we could take the raw percentages as probabilities of superiority and use them to answer this question. However, it is unlikely that the relationship between mark difference and probability of superiority will be a smooth progression. More likely, we will have something like the pattern evident in **Table 1**.

It is not clear from this whether the 50% probability of the paper 2 script being judged superior is at a mark difference of 0 marks, or between 2 (40%) and 3 marks (71%).

To overcome this issue, the simplified pairs method uses a logistic regression to generate a smoothed relationship between the mark difference and the probability of the paper 2 script being judged superior. In this type of model, for the ith pair of scripts judged by the jth judge we denote the difference between the mark awarded to the paper 2 script and that awarded to the paper 1 script as d_{ij}. We set $y_{\mathrm{ij}} = 0$ if the judge selects the paper 1 script as superior and $y_{\mathrm{ij}} = 1$ if they select the paper 2 script. The relationship between y_{ij} and d_{ij} is then modelled using the usual logistic regression equation:

$$P(y_{\mathrm{ij}} = 1) = \{1 + \exp(-(\beta_0 + \beta_1 d_{\mathrm{ij}}))\}^{-1}$$

From this equation we need to find the value of d_{ij} such that $P(y_{\mathrm{ij}} = 1) = 0.5$. This will give us the mark difference associated with a probability of 0.50 that the paper 2 script will be judged superior. If we denote the estimated coefficients in the logistic regression model as $\widehat{\beta_0}$ and $\widehat{\beta_1}$, then after some re-arranging the estimated difference in difficulty for a probability of 0.50 is:

$$\widehat{d} = \frac{-\widehat{\beta_0}}{\widehat{\beta_1}}$$

This difference can then be compared with the empirical difference in difficulty between the two papers.

As this method only requires each script to appear in one paired comparison, it has a notable advantage over alternative CJ methods, which generally require that each script appears in many comparisons. The results of these comparisons are then combined and analysed using a statistical model, such as the Bradley-Terry model (Bradley and Terry, 1952, p. 325). Simplified pairs does not require this step, meaning that a much greater number of scripts can be included in a simplified pairs study compared to Bradley-Terry methods, without the judges having to spend any more time on making judgements.

It is also important to consider the design of CJ exercises. There are several aspects to this, including the choice of papers, the number of scripts and judges, the range of marks that the scripts will cover, and the instructions to the judges. These are outlined in the following sections.

Choice of Papers

The first step in each study involved selecting two assessments to be used for the analysis. We used GCSE (General Certificate of Secondary Education) assessments for both exercises, which are typically sat by students at the age of 16 in England.

For the mathematics exercise, we created the assessments by splitting a single 100-mark GCSE Mathematics exam component into two 50-mark examinations ("half-length assessments"). The original full-length assessment for analysis was chosen as it was taken by a large sample of students, which meant that we could undertake a formal statistical equating, to use as a comparator to the results from the simplified pairs.

Further details on the two half-length assessments are displayed in **Table 2**, and some example questions are listed in Appendix B (see **Supplementary Material**). Each half-length assessment contained 10 questions worth a total of 50 marks. The mean scores of each question were calculated based on the responses of all 16,345 candidates and are also displayed. As can be seen, the total of these mean question scores indicates that Half 2 was roughly 5 marks harder than Half 1.

For science, the exam papers we chose were the foundation tier chemistry papers from the OCR Combined Science A GCSE qualification. Two papers, named component 03 and component 04, were used. Example questions are listed in Appendix B (see **Supplementary Material**). As with the Mathematics paper, these papers were chosen partly because they were taken by many students. An additional reason for choosing these papers was that the mean mark was higher on component 03 by around 9% of the maximum. One aim of this research was to see if examiners could make allowances in their judgements for differences in paper difficulty, and this seemed like a reasonable level of difference (i.e., challenging, but not impossible).

TABLE 2 | Details of questions included in each half-length assessment in the mathematics study.

Question	Mean question scores		Max question scores	
	Half 1	Half 2	Half 1	Half 2
Q1	3.34		4	
Q2	0.85		1	
Q3		4.32		7
Q4	7.86		9	
Q5		4.44		6
Q6	3.40		6	
Q7		3.69		6
Q8	2.02		5	
Q9	2.40		6	
Q10		1.89		5
Q11		1.13		4
Q12		3.15		4
Q13	4.30		5	
Q14		1.92		3
Q15	2.74		7	
Q16		1.32		3
Q17	1.22		3	
Q18		1.74		6
Q19	2.05		4	
Q20		1.64		6
Total	**30.19**	**25.22**	**50**	**50**

Both science papers had a maximum mark of 60 and were each worth 1/6th of the whole qualification for foundation tier candidates. However, it is worth noting that the science papers did not cover the same content. This contrasts with the situation in a typical standard maintaining exercise (such as awarding), when the two papers being compared are based on mostly the same content. It also contrasts with the mathematics exercise described here and with previous trials of the simplified pairs method (e.g., Benton et al., 2020). Therefore, the examiners in this exercise may have found the task harder than a similar task undertaken to assist with awarding.

Choice of Scripts

In both exercises, exam scripts were randomly selected for the simplified pairs comparison exercise, 300 from each paper (or half paper in the case of mathematics). As the exercises were independent of one another, this means 300 scripts were selected in mathematics and 300 in science; these were all different students. For mathematics, different samples of students were used to provide script images for the Half 1 assessments and for the Half 2 assessments.

In standard maintaining exercises we are interested in determining changes in difficulty across the whole mark range (or at least the mark range encompassing all the grade boundaries). Therefore, in CJ studies in this context it is important to ensure that the paired comparisons include scripts with a wide range of marks in both papers.

For each half-length assessment in mathematics, scripts with scores between 10 and 45 (out of 50) on the relevant half were selected, with an approximately uniform distribution of marks within this range. Scripts from each half were randomly assigned to pairs subject to the restriction that the raw scores of each half-script within a pair had to be within 15 marks of one another.

For science, the intention was that the spread of scripts across the mark range was the same for both components (an approximately uniform distribution from 20% to 90% of maximum marks). However, due to a small error in the code used to select the scripts, the range for component 04 was actually from 13% to 90% of the maximum mark. This contributed to the fact that the average score for the scripts selected for component 04 was around 4.5 marks lower than for component 03. This error was not picked up until after the exercise was complete. It will not have had any effect on the statistical analysis, as there were still many comparisons made across a broad range of mark differences. However, it is possible that it had a psychological impact on the examiners (who might have expected a more even distribution of mark differences).

The range of marks on the scripts was 12 to 53 on component 03 and 8 to 48[4] on component 04. The scripts were randomly assigned to pairs. For some pairs, the component 03 script had a higher mark and for some the component 04 script did. Some pairs had a very large difference in marks, whilst others had a difference of zero marks.

Examiner Instructions

Six experienced examiners were recruited to take part in each exercise – i.e., six for mathematics and six for science. However, in the GCSE Science exercise, one judge subsequently dropped out due to other commitments. Each examiner was given 50 pairs of scripts (half-scripts for mathematics) to compare (on-screen), and they were asked to determine 'Which script is better, based on overall quality?'.

The examiners were given additional guidance explaining that this involved making a holistic judgement of the quality of the scripts, using whatever method they wished, to choose the better one. They were also told that they should use their professional judgement to allow for differences in the relative difficulty of each test. In advance of the task, the examiners were provided with the exam papers and mark schemes and asked to re-familiarise themselves with both. Beyond this, there was intentionally no specific training provided, as the rationale of CJ is for examiners to use their professional judgement to make holistic judgements.

All judgements were made on-screen using the Cambridge Assessment Comparative Judgement Tool[5]. No marks or other annotations were visible to the judges on any of the scripts.

Figure 1 shows further details on the design of the science CJ task (and some of the results). Each numbered 'point' on the figure represents one pair of scripts, with the number indicating the examiner making the judgement. The horizontal axis shows the mark given to the script from component 03 and the vertical axis the mark given to the script from component 04. Blue indicates that the script from component 03 was judged superior, and red that the component 04 script was judged superior. The diagonal line is a line of equality between the two marks, so that points below the line indicate a pair where the script from component 03 had a higher raw mark than the component 04 script. Unsurprisingly, blue points were more likely to be below the line and red points more likely to be above the line. More detailed analysis of the relationship between assessment scores and judges' decisions will be shown later in this article.

Follow-Up Surveys of Judges

The follow-up surveys provided data to address Research Question 2 – "How do judges make comparative judgements of students work from exam papers comprising many highly structured tasks, and what validity implications does understanding their processes have?" After both studies, the judges who had taken part were invited to take part in short surveys to inform the researchers about their experiences of the task and about how they thought they made their judgements. The surveys were administered to judges *via* SurveyMonkey$^{\text{TM}}$ a short time after they had finished their judgements and took approximately 10 min to complete. The two surveys were slightly different in their questions, but similar enough for answers to be compared here. Insights from the surveys, especially those that relate to the validity of the exercise, are discussed in what follows.

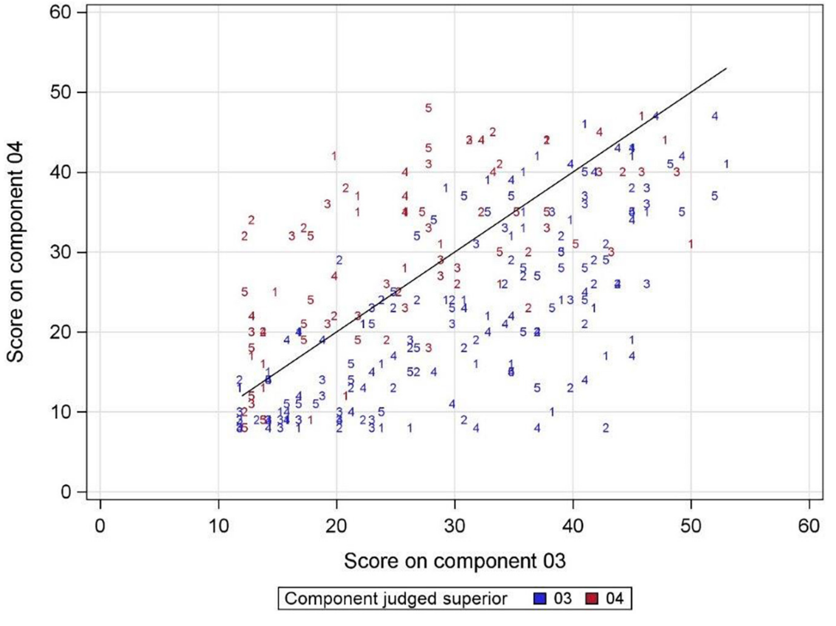


FIGURE 1 | The design of the simplified pairs study. The locations of the points show which scores on science component 03 were paired with which scores on component 04 and the numbers indicate which examiner made the judgement. The black line is a line of equality, rather than a regression line.

[4] This is only 80% of the maximum mark, but there was only one candidate who achieved a mark of more than 48 on this component.

[5] https://cjscaling.cambridgeassessment.org.uk/

These surveys were developed by the researchers, based on those used in earlier exercises within the wider series of studies, with the precise wording and focus of questions being arrived at as a result of an iterative process. The questions are listed in Appendix A (see **Supplementary Material**), except where they are not directly related to the subject of this article. A combination of open questions and closed-response questions using five-point Likert scales were utilised. Surveys were used in order to get responses rapidly, in order that insights could be acted upon within the wider series of studies in terms of future exercise design.

Data from open questions were analysed using both *a priori* and inductive methods. *A priori* themes were predicted from previous experience and literature (on issues including approaches to judging where students had uneven performance across a paper). Inductive methods revealed new themes (such as differences in approach between mathematics judges and science judges).

RESULTS

Overall Difference in Difficulty

Here, and in Section "Simplified Pairs Results," we present results that answer RQ1. Firstly, the overall difference in difficulty between the two assessments in each exercise is shown. **Table 3** presents the results of a mean equating between the two half-length assessments in the Mathematics GCSE, which demonstrates the empirical difference in difficulty between the two half papers. **Table 4** presents the results of the equating between component 03 and component 04 in the Science GCSE. These were based on the scores of all students taking the component(s), not just those included in the CJ exercise. The tables show that for the mathematics exercise, Half 2 was about 5 marks harder than Half 1 and for science, component 04 was about 5 and a half marks harder than component 03.

Simplified Pairs Results

Next, we present an estimate of the overall difference in difficulty between components (or half papers) using the results of the simplified pairs exercise. **Figures 2**, **3** plot the proportion of paired comparisons where the script from Half 2 (mathematics) or component 04 (science) was judged superior, against the mark difference between each pair of scripts. Larger points depict mark differences with more judgements made. As can be seen, the proportion of pairs where Half 2 (or component 04) is deemed superior tends to increase with the extent to which the mark on the Half 2 (or component 04) script exceeds the mark on the Half 1 (component 03) script.

The formal analysis within a simplified pairs study was done using logistic regression[6]. This is represented by the solid red line in **Figures 2**, **3** which smoothly captures the relationship between mark differences and the probability of a Half 2 (or component 04) script being judged superior. The main purpose of this analysis is to identify the mark difference where this fitted curve crosses the 0.5 probability. For mathematics, this happens at a mark difference of −3.4. This implies that a Half 2 script will tend to be judged superior to a Half 1 script wherever the mark difference exceeds −3.4. In other words, based on expert judgement we infer that Half 2 was 3.4 marks harder than Half 1.

A 95 per cent confidence interval for this value (the dashed vertical lines) indicates that the judged difference in difficulty was between −2.4 and −4.3 marks. It should be noted that the size of this confidence interval, of essentially plus or minus a single mark, was very narrow compared to previous published examples of both simplified pairs (Benton et al., 2020, p. 19) or other kinds of CJ in awarding (Curcin et al., 2019, p. 11). This was because the relationship between mark differences and judges' decisions depicted in **Figure 2** was much stronger than in many previous applications, leading to increased precision.

The estimated difference based on expert judgement (*via* simplified pairs) fell a little short of the true difference at only 3.4 marks. Furthermore, the confidence interval for the simplified pairs estimate did not overlap with the empirical difference. This indicates that we cannot dismiss the differences in results from mean equating and simplified pairs as being purely due to sampling error. Nonetheless, the exercise correctly identified the direction of difference in difficulty and the estimate was close to the correct answer.

For science, the curve crosses the 0.5 probability at a mark difference of 1.3 marks, which indicates that, according to examiner judgement, component 04 was easier by just over 1 mark. The 95% confidence interval was between −0.7 and 3.3 marks. As this range includes zero, we cannot be sure, from

TABLE 3 | Results from mean equating of the actual scores of pupils taking the two half papers (mathematics).

	Half 1	Half 2
Number of students	16,345	16,345
Mean score	30.19	25.22
SD score	9.78	9.71
Difference in means (Half 2 – Half 1)	−4.96	
SE of difference in means	0.04	
Confidence interval for difference in means	[−5.04, −4.88]	

TABLE 4 | Results from mean equating of the actual scores of pupils taking the two components (science).

	Component 03	Component 04
Number of students	10,043	10,043
Mean score	24.13	18.72
SD score	8.48	8.82
Difference in means (component 04 – component 03)	−5.41	
SE of difference in means	0.05	
Confidence interval for difference in means	[−5.51, −5.31]	

[6] For more details on this method, see Benton et al. (2020).

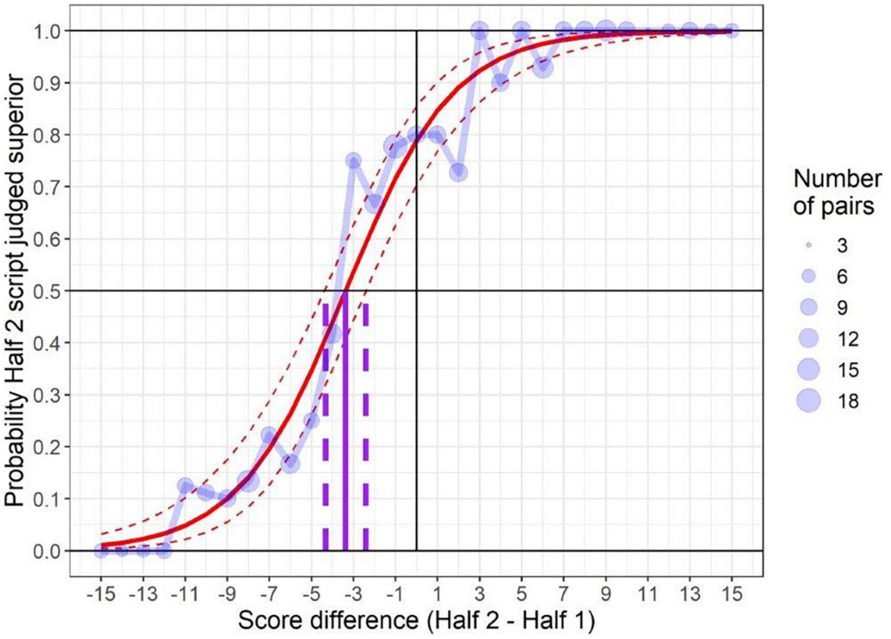


FIGURE 2 | Graphical depiction of the results of using simplified pairs to gauge the relative difficulty of two assessment versions (mathematics).

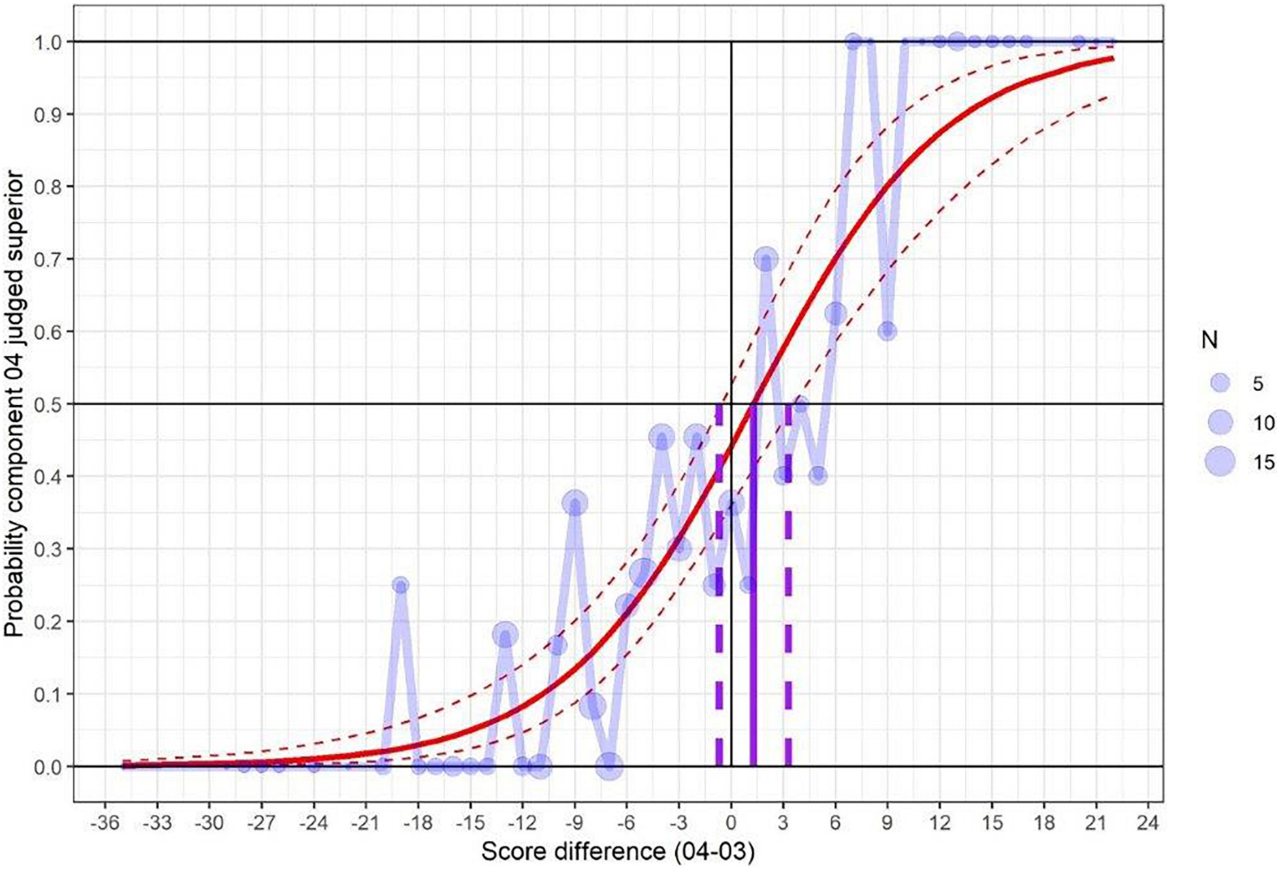


FIGURE 3 | Graphical depiction of the results of using simplified pairs to gauge the relative difficulty of two assessment versions (science).

judges' decisions, that there was any difference in difficulty between the two papers.

This result contrasts with the empirical difference in difficulty according to the marks (**Table 4**), which revealed that component 04 was around 5.4 marks harder. This result will be discussed further in the "Discussion" section of this report.

Judge Fit

Tables 5, **6** show some statistics on the judge fit for both exercises, and how long the judges took on average to make their judgements. A visual depiction of how the fitted logistic curves differed between judges is shown in **Figures 4**, **5**.

In both exercises, each judge displayed strong point biserial correlations between the differences in marks for the half-scripts (or components) being compared and the decision they made about which was superior. The range of point biserials in the two exercises (between 0.63 to 0.82 in mathematics and between 0.42 and 0.67 in science) compared well with the range shown in studies of more subjectively marked subjects such as English Literature, as explored in Benton et al. (2020, p. 22), where judges' point-biserial correlations were between 0.33 and 0.62. This reiterates the strong relationship between mark differences and judges' decisions in both exercises considered in this article.

In the GCSE Mathematics exercise, all six judges selected Half 2 as being superior more than 50 per cent of the time and, similarly, each of the logistic curves for separate judges intersects the 0.5 probability line at mark differences below zero. This indicates a unanimous suggestion across judges that Half 2 was a harder assessment than Half 1.

In GCSE Science, the picture was more mixed. Results from judges 1, 4 and 5 would suggest that component 04 was easier (by between 1 and 4 marks), whereas for judge 3, component 03 came out as easier (by about 2.5 marks). For judge 2, there was almost no difference in difficulty. This lack of agreement about which paper was easier contrasts with previous research, where judges agreed unanimously about which paper was harder. However, the differences (in terms of paper difficulty) between judges in the current exercise were not that large and were similar to those found in the previous research.

Furthermore, although four out of the five judges had similar shaped curves, the results from judge 4 were somewhat different. This judge had a much steeper curve, pointing to a more ordered set of decisions about which script was superior. We looked more closely at the decisions of this judge and found that the relationship between mark difference and decision was almost perfect, with only one judgement out of order: the examiner judged component 04 to be superior for all mark differences of 4 or more, and judged component 03 to be superior for all mark differences of 3 or less (with one exception). This suggests that this judge was actually remarking the scripts and then basing their decision of superiority on a pre-conceived idea about the difference in difficulty between the two components. Interestingly, that pre-conception was that component 04 was easier by about 4 marks, which was very different from the empirical difference (component 04 harder by 5.5 marks).

Tables 5, **6** also show judge fit calculated using INFIT and OUTFIT[7] (see Wright and Masters, 1990). For mathematics, none of the values are high enough (or low enough) to warrant serious concern over any of the judges. The highest values occur for the two judges (judges 1 and 6) with logistic curves (**Figure 4**) that suggest the smallest estimated difference in the difficulty of the tests. For science, Judge 4 stands out as having particularly low values of INFIT (0.50) and OUTFIT (0.33), which suggests over-fitting of the data to the model, consistent with this judges' apparent tendency to re-mark. However, since decisions within the exercise are to some extent a matter of opinion (see Benton et al., 2020, p. 10) we tend to prioritise information from point biserials over judge "fit".

The median time required per judgement was between 2.2 and 5.6 min for mathematics and between 4.9 and 6.7 min for science. There was quite a strong negative relationship in science between the median time and the point biserial correlation, with longer median time associated with a lower correlation. This suggests that some of the examiners may have found it a more challenging task, and this meant they were both slower and less accurate.

Equating Across the Score Range

In **Tables 3**, **4** we presented the overall empirical difference in difficulty between the two components (or half papers), using mean equating. We now extend this further by equating these across the full mark range. For this we used equipercentile equating, which generated an equivalent mark on Half 2 (or component 04) for each mark on Half 1 (component 03). This was done using the R package equate (Albano, 2016). The results of the equating were then compared with the equivalent

[7] INFIT and OUTFIT indicate how closely the empirical data fits the modelled data (from the logistic regression model) for each judge. Values larger than 1 indicate un-modelled noise, values lower than 1 indicate over fit of the data to the model.

TABLE 5 | Judge fit and speed for each of the six judges (mathematics).

Judge	No. of pairs	Proportion with Half 2 selected	INFIT	OUTFIT	Point biserial correlation between difference in marks and selecting half 2	Median time per judgement (minutes)
1	50	0.62	1.53	1.58	0.73	3.5
2	50	0.56	0.58	0.26	0.82	5.1
3	50	0.70	0.73	0.34	0.77	2.2
4	50	0.72	1.10	0.74	0.63	4.2
5	50	0.58	0.68	0.43	0.82	5.6
6	50	0.58	1.34	1.43	0.63	4.2

TABLE 6 | Judge fit and speed for each of the five judges (science).

Judge	No. of pairs	Proportion with component 04 selected	INFIT	OUTFIT	Point biserial correlation between mark difference and selecting component 04	Median time per judgement (minutes)
1	50	0.34	1.28	1.75	0.42	6.7
2	50	0.36	1.00	0.85	0.64	4.9
3	50	0.42	1.13	1.21	0.52	5.6
4	50	0.18	0.50	0.33	0.67	5.1
5	50	0.34	1.04	0.80	0.57	4.9

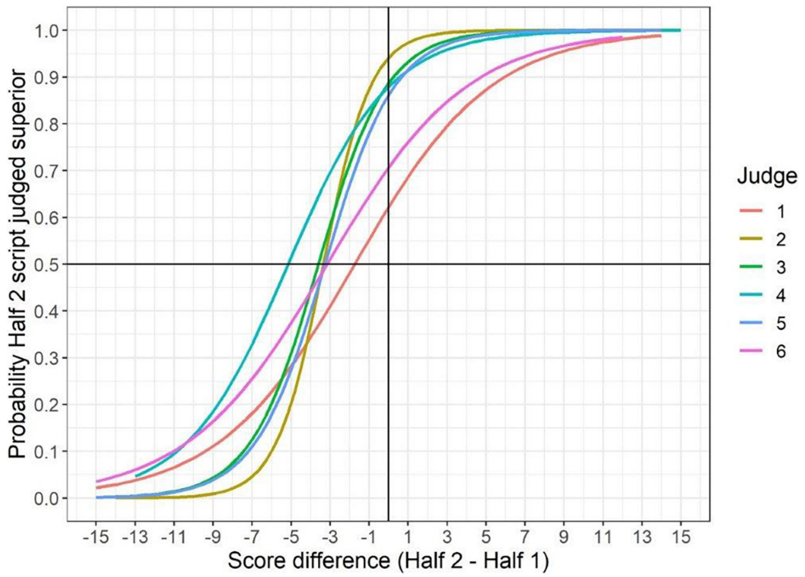


FIGURE 4 | The relationship between differences in marks for a pair and the likelihood of selecting half 2 as superior (by judge).

marks generated by the logistic regression results from the simplified pairs exercise.

Figure 6 (mathematics) and **Figure 7** (science) present the results of this analysis, with the red lines showing the results according to the equating, and green lines the results according to the CJ exercise. The dashed lines represent 95% confidence intervals for the equivalent marks. For reference, the graph also includes a straight diagonal line of equality.

In mathematics, the results from empirical equating (the red line) confirm that Half 2 was harder than Half 1. This difference in difficulty is particularly visible for marks between 25 and 45 marks on Half 1. A similar pattern is also visible from the results of simplified pairs (the blue line) indicating a reasonable level of agreement between the two techniques.

According to the empirical equating for science, component 04 was harder than component 03 across the whole mark range, with the difference steadily increasing between marks of 0 and 20 on component 03 (up to a maximum of 6.2 marks). Above this mark, the difference fell steadily up to a mark of 45, above which there were only a few candidates so there was less certainty about the equivalent mark on component 04. No candidates achieved a mark higher than 53 on component 03. The equivalent marks according to the results of the CJ exercise were very different, varying between 0 and 1 mark easier for component 04. The confidence intervals for these marks were also substantially wider than those generated by the equating. Only at the very top of the mark range does the confidence interval for the simplified pairs method encompass the estimate from equating.

Insights From Surveys of Judges

This section provides insights from the surveys, which help to answer RQ2. Results are presented here in a narrative fashion, in order to explore findings in more detail. Answers to both Likert-scale and open questions are integrated into what follows, while descriptive statistics, as they would offer little insight, are not presented in tabular form.

The judges were asked how straightforward they found the process of making a holistic judgement of script quality. One science judge responded that this was 'very straightforward' and three considered it 'somewhat straightforward'. The remaining judge said they were not sure and admitted to 'counting points' to start with. In the mathematics survey, five out of the six judges said it was at least somewhat straightforward, with two of them believing it to be entirely straightforward. The sixth considered the process to be 'not very straightforward', noting that given that mathematics papers contain lots of questions of differing demand, making a holistic judgement of mathematics papers was in their view very difficult. They highlighted that it would be

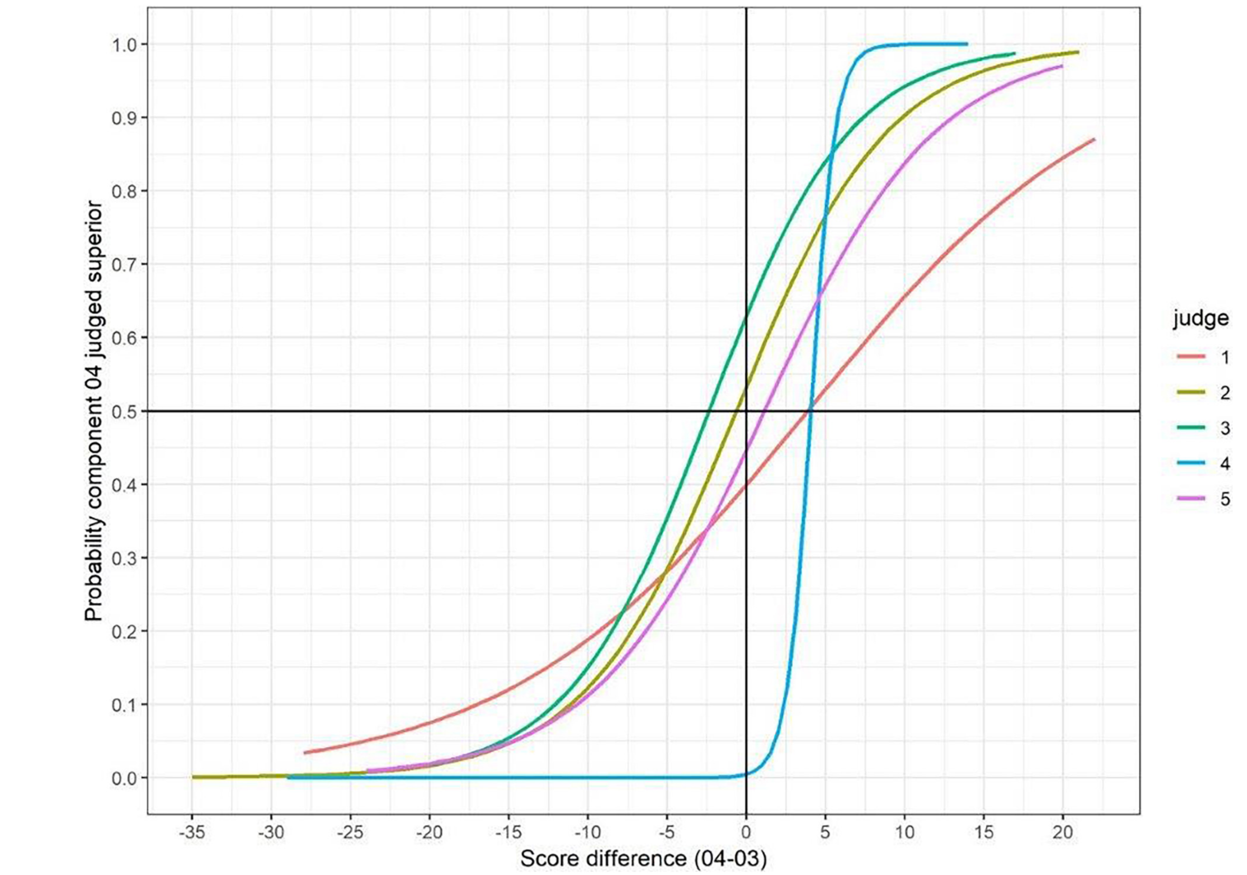


FIGURE 5 | The relationship between differences in marks for a pair and the likelihood of selecting component 04 as superior (by judge).

easier to compare two responses to the same question, or two sets of questions of the same standard. Another mathematics judge, who had difficulties making holistic judgements initially, nonetheless said that this grew easier over time.

Across the two exercises, the judges suggested different specific processes for making their judgements. When asked how they believed that they made their judgements, all science judges said that they looked at the answers to key questions. However, four out of the five judges also said that at times they needed to look at the number of correct responses in each paper. This was particularly the case for lower quality scripts, where there were often no responses to the questions with more marks. One science judge also mentioned that they ignored the responses to multiple-choice questions, an interesting finding potentially highlighting the extent to which CJ methods can be considered more applicable to longer-response items. Meanwhile, in mathematics, some judges focused on the number of answers correct, while others attempted to match questions on each half of the paper by either (a) their perceived difficulty or (b) the skills required to answer them, and then tried to determine which script was superior. Candidates' working was considered by two mathematics judges to be a significant discriminator, while another highlighted communication. Many of the judges said that they used many of these different processes at the same time.

In both surveys, the judges were asked directly which of the two papers they were judging they believed to have the more difficult questions. Three out of the five science judges thought that the two different exam papers were of the same level of demand, whilst one thought (correctly) that component 04 was more demanding and one judge was unsure. The judge who was unsure put this uncertainty down to the fact that 'the scripts were rarely assessing the same assessment objectives with comparably scored questions, so the level of demand varied'. Encouragingly, four out of the six mathematics judges correctly identified Half 2 as the more difficult of the two half-length assessments, while another saw the two halves as very similar in difficulty. The sixth noted only that one half was more difficult than the other but did not specify which. Here it can be seen that the outcomes of the two exercises respond to the view that examiners find it difficult to make judgements about overall paper difficulty (e.g., Good and Cresswell, 1988, p. 278) in diverse ways – that is, the mathematics judges were able to correctly approximate the empirical difference in difficulty of the two half-scripts, while the majority of the science judges could not do this.

Judges differed in their views as to whether questions worth more marks are invariably better discriminators of candidate quality. Those mathematics judges agreeing to this contention highlighted that high-tariff questions allow for problem-solving

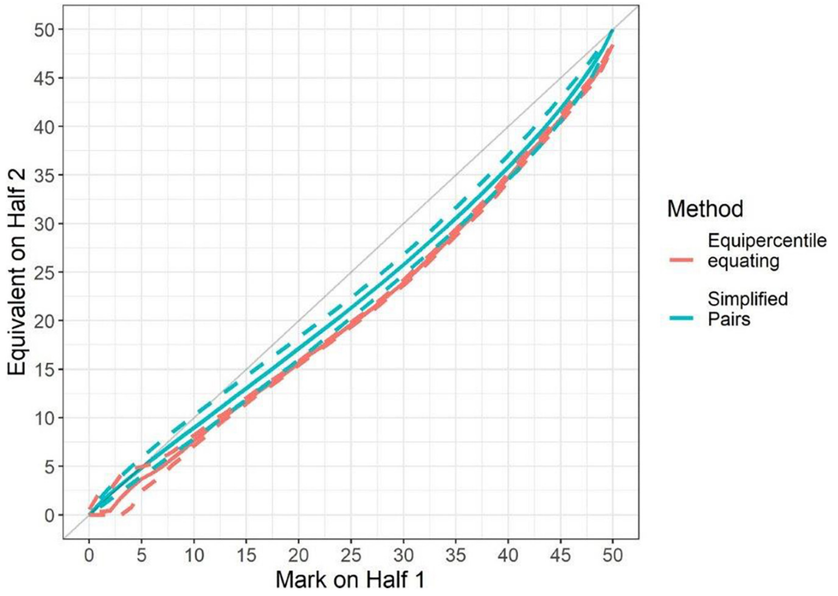


FIGURE 6 | Results of both equating and simplified pairs across the score range (mathematics).

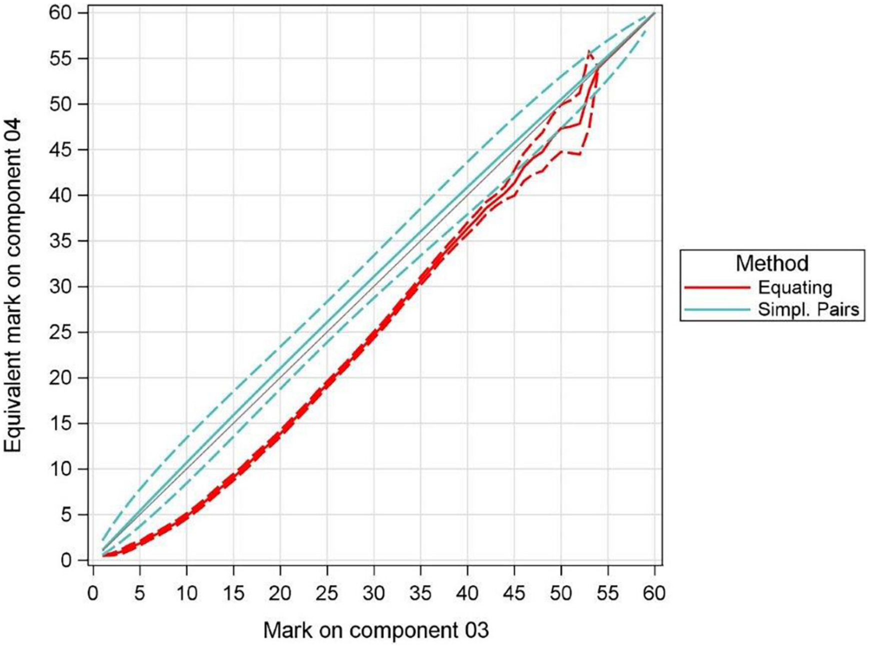


FIGURE 7 | Results of both equating and simplified pairs across the score range (science).

skills to be evidenced and are often of greater complexity, while those opposed noted that some high-tariff questions can be quite routine, and can be prepared for, while communication issues can be more revealing in low-tariff questions. There was also disagreement among science judges on the same theme, with two agreeing somewhat, two disagreeing somewhat and one neither agreeing nor disagreeing. One science judge who agreed said that the higher tariff questions required answers involving explaining or analysing. Of those who disagreed, one stated that the statement was not true for weaker candidates, because they achieved fewer marks on the higher tariff questions. This is an interesting response, because at first glance it sounds like a definition of a discriminating question. Perhaps they were suggesting that most low ability candidates would get zero on the higher tariff questions, meaning that it would be impossible to discriminate between them. Another judge thought that the six-mark questions test writing ability as much as chemistry ability.

There was more agreement than disagreement in both studies with the contention that certain types of questions were better discriminators than others. All science judges agreed (three 'entirely' and two 'somewhat') that some types of questions were better discriminators. Only three of the science judges expanded on their response, with general agreement that questions requiring interpretation, application or explanation responses were better discriminators. One science judge elaborated further by saying that this was the case for high performing candidates, but that for weak candidates the 'recall' questions were better discriminators. Mathematics judges had similar responses. An additional question asked more explicitly about what types of questions were better discriminators. The most selected responses in the science context (selected by four judges each) were 'questions testing application of knowledge' and 'questions involving analysis of information and ideas'. Mathematics judges offered varied opinions in response to the same question, including multi-part questions, knowledge and understanding questions, and data analysis.

There was also no strong agreement between mathematics judges as to whether they believed they did consistently focus on particular types of questions in their judgements, some suggesting that unstructured questions might be a useful tiebreaker but others attempting to make holistic judgements based on all types of questions across the paper. Only three of the science judges said that they focused on certain question types when making their judgements, two of whom said that they focused on 'questions testing knowledge and understanding'. The remaining science judge selected 'other' as their response, but their explanation suggested that they too focused on 'questions testing application of knowledge', alongside 'levels of response questions'. One judge said that the reason they did not focus on certain question types was that they were asked to look at the whole script when making their judgements. Overall, though, it is reassuring that the judges were mainly focusing on the same question types when making their judgements, because it suggests a degree of consistency in their method.

Many mathematics judges described difficulties in making judgements of pairs where a candidate's response to one half-script was better in one sense, but worse in another sense, than the other candidate's response to the other half-script in the pair. For example, one mathematics judge noted an example where one candidate performed more strongly on trigonometry, but less well on algebra, while another indicated an example where one candidate answered every question, though not entirely correctly, while the other produced correct solutions to about half the questions. Most judges suggested that the tiebreaker for them in such cases was performance on the higher tariff, "harder" questions towards the end of the paper. One of the challenges here is that it is by no means clear what the correct tiebreaker "should" be, in this context. It is worth noting that this same issue arises even when making comparisons within the same test (Bramley, 2012, p. 24). As such we cannot expect holistic judgements of quality to match the mark scale exactly.

Finally, in their survey, two mathematics judges indicated a belief that comparative judgement methods might work less well for mathematics than subjects involving longer, more discursive answers such as English or history. This relates back to the general question that underpins this article – does CJ struggle in relation to papers comprised mainly of highly objective, short-answer questions (as mathematics and science papers typically are), because of the difficulty for judges of knowing how to sum the many different small bits of evidence of candidate quality presented in each item (taking into account the variance in item difficulty) in coming to a holistic judgement. While the outcomes in these two exercises lead to an equivocal finding in relation to this question, what is perhaps likely to be less equivocal is the attitude of judges to whether they think they can do what is required. For comparative judgement to be operationalised, the support of those intended to be used as judges would be vital.

Principally, the concern here lies in the fact that, in many mathematics assessments, achieving the right answer the most times is the main objective (it "boils down to right or wrong", according to one judge). This was also highlighted by judges who noted that it can be difficult to avoid simply re-marking the scripts. It was suggested that the need to bear in mind many small judgements of superiority (of candidates' performance on questions testing different skills, for example) and then combine them into one overall judgement, for example, leads to more cognitive load and a more tiring task than marking, again suggesting that it may perhaps be difficult to establish judge support for the greater use of comparative judgement in the future.

On the other hand, most of the judges had never taken part in a comparative judgement exercise before and their experiences varied. More experienced judges might have been more consistently supportive. Moreover, it should be acknowledged that the surveyed judges did mostly say that the process was straightforward (at least once they had got into it) – implying that, as is often the case with complex changes to processes, while there might be hesitation initially, eventually this would give way to acceptance and then confidence.

The information revealed in the judge surveys helps us gain a better understanding of the ways in which the judges in both studies made their judgements. It also offers some clarity as to issues around what parts of the papers the judges were attending to, such as the relative importance of higher- and lower-tariff questions to judges' decisions and the comparative significance of diverse types of questions in demonstrating candidate quality.

However, what is perhaps most striking about the survey outcomes is that (a) there is no consistency across judges in the same survey about what they regard as important and (b) a difference in what is regarded as important between mathematics judges (considered collectively) and science judges (considered collectively) that might explain the difference in outcomes between the two studies is not evident. In other words, it is not the case that, for instance, mathematics judges thought that they were clearly much more capable than science judges at determining the difficulty of questions, or that science judges were clearly not as good at deciding which questions to focus on. This means it is not easy to explain why one of the exercises "succeeded" and the other did not, at least by reference to the judges' processes.

DISCUSSION

The findings of the two studies discussed above offer several points for further discussion relating to the appropriateness, accuracy, and validity of the use of comparative judgement in subjects with highly structured papers. In Section "Overall Outcomes," we discuss the overall outcomes of the exercises, in order to address RQ1. Then, in Section "How Judges Judge," insights from the surveys are discussed to explore some of the validity issues relevant to RQ2.

Overall Outcomes

Firstly, the overall outcomes of the two exercises are discussed, in order to answer RQ1. It is notable that the two studies, despite being nearly identical in structure, resulted in somewhat different outcomes. In the mathematics study, the results of the simplified pairs exercise meant that, based on expert judgement, we can infer that the Half 2 paper was 3.4 marks harder than Half 1. This was about 1.6 marks away from the empirical difference estimated from statistical equating, where Half 2 was 5 marks harder than Half 1. Judges unanimously agreed that Half 2 was a harder paper than Half 1, suggesting that it is possible for judges to make determinations of test difficulty. This appears to at least somewhat allay Baird's (2007, p. 142) concern that examiners cannot compensate for the differing demands of question papers from year to year. In line with Benton et al. (2020, p. 21), we suggest here that our judges were able to appropriately make allowances for paper difficulty in this exercise at least.

However, in the science study there was no consistency between judges as to which component was harder. The fact that there is a distinction between the studies is a somewhat discouraging finding in terms of the consistency of CJ. This lack of consistency between judges is despite the fact that in the science study, the empirical difference between the papers, as estimated from statistical equating, was 5.4 marks, with component 04 harder than component 03. Results from three judges suggest that component 04 was easier (by between 1 and 4 marks), whereas for another judge, component 03 came out as easier (by about 2.5 marks) and for a final judge, the two papers were almost equal in difficulty. Overall, these judgements amounted to component 04 being viewed as about 1 mark easier than component 3, but this was not statistically significantly different from zero (no difference). It is important to note that, despite these differences, the variability of results from different judges in terms of their assessments of paper difficulty were not that large as a percentage of the maximum mark on the paper.

On the other hand, the range of point-biserial correlations between mark difference and the likelihood of selecting the second of the two papers as the harder was between 0.63 and 0.82 in mathematics and 0.42 and 0.67 in science. This means that the judges were both mostly consistent with each other, in terms of working out which paper was harder, and their judgements were mostly accurate. The range of point-biserials here is not far from the range demonstrated in CJ exercises concerning more subjectively marked subjects where papers are constructed from a smaller number of less structured extended tasks, such as English Literature. See, for example, Benton et al. (2020, p. 22), where the range was between 0.33 and 0.62. Both exercises – mathematics and science – therefore demonstrate a strong relationship between how far apart the papers in any pair were in terms of marks, and the judges' likelihood of correctly determining which was superior. The fact that these ranges were similar to those evident for papers with more extended tasks is encouraging. However, if judges are not capable on the whole of correctly determining which of the papers was harder, as was the case in the GCSE Science exercise, the consistency of their judgements matters less – in other words, are they just reliably incorrect?

Judge fit (consistency) also has some value for the validity of the exercise, though this is of less significance in terms of illustrating the accuracy of the exercise (RQ1) than the point-biserial correlation between the judge's decision and the mark difference between the papers they were judging. Moreover, few judges stood out in terms of their INFIT and OUTFIT values. On the other hand, it could be suggested in relation to RQ2, that, where judges misfit the model, this could be because there were re-marking rather than making holistic judgements. The activity of re-marking is related to the structure of the items. Re-marking is more likely in a structured question paper than a paper requiring extended responses. This highlights the need for further work to address the question about the meaning of holistic judgement in CJ and its relationship to processes such as re-marking; this conversation has also been contributed to by Leech and Chambers (2022).

How Judges Judge

Both exercises also offer interesting insights in relation to the processes that judges used to make their judgements, and how they found the exercises, which can help to answer RQ2. The fact that a majority of judges in both studies considered it at least "somewhat straightforward" to make holistic judgements is encouraging, although at least one mathematics judge offered a contrary view, arguing that given that mathematics papers contain lots of small questions of differing demand, a holistic judgement was difficult to arrive at. However, this was a minority view. These findings accord with those of earlier studies involving papers with more extended tasks (e.g., Greatorex, 2007; Black, 2008; and Jones et al., 2015), suggesting that there is nothing specific about the fact these papers had highly structured tasks that meant judges felt it was less straightforward to judge them holistically.

However, the cognitive load put on judges who have to sum up many different small pieces of evidence, while taking appropriate account of the difference in difficulty of the papers overall, is clearly substantial. This echoes the findings of Verhavert et al. (2019) who found that the structure of a task impacts the complexity of decisions made by judges. Moreover, there are significant commonalities with the work of Leech and Chambers (2022), who found that in more structured papers many judges were making judgements that were, at best, partially holistic. We can therefore see that this problem is more evident in papers made up of highly structured tasks as is typical of United Kingdom exam board papers in mathematics and science. Finally, whether judges can correctly assess and take account of the difficulty of papers (as questioned by Good and Cresswell, 1988, p. 278) is something that these studies provide only ambiguous evidence on.

Processes

The survey findings from these studies are generally similar to those relating to earlier CJ exercises (concerning papers with more extended tasks) in the insights they provide about the processes that judges use. That is, that different processes are used by judges, with many judges utilising many of the processes at the same time, but outcomes are generally consistent with one another. For example, all science judges looked at answers to key questions, as was the case in the study by Greatorex et al. (2008, pp. 4–5); and most at the number of correct responses. The fact that judges across both studies used a variety of different processes, and yet were generally consistent with one another (in the same study), suggests that the ability to make a holistic judgement of script quality is not necessarily directly related to the specific process used to make that judgement.

In one respect, this is a good thing, since it is generally acknowledged as a strength of marking that it involves processes that are relatively consistent across markers, and so the fact that outcomes (if not processes) are consistent in CJ is encouraging. However, from a public confidence viewpoint, does the variation in judgemental and discriminatory processes used by CJ judges have the potential to cause disquiet? Current marking and awarding processes value standardisation and transparency, which CJ does not in the same way. The issue of the different approaches used by different judges may be of concern, particularly in relation to the ability to maintain an audit of how decisions have been made. The work of Chambers and Cunningham (2022) on other aspects of decision-making processes in CJ is also important in this regard.

Questions Attended to

A follow-on issue from that of process is that of which items in the papers judges most frequently attended to. Judges did not agree about whether higher-tariff questions were more useful *in general* for their judgements; instead, which questions were more helpful depended on their type and what skills they were testing. Overall, though, it does appear that judges were generally focusing their attention on certain questions. Generally, the same kinds of questions were focused on in each study. Some subject-specific issues arose, including the key role judges saw for candidates showing their working in mathematics, and the idea of skills application and analysis in science, indicating the many different concerns at play in the assessment of candidates in different subjects.

Other causes of challenging decisions include where the writer of one script in a pair was better at one skill or in one section of the paper but the writer of the other was better at another skill or section, and each is important; a general instruction to make a holistic judgement may not be clear enough to guide judges in these cases. A variety of heuristics seemed to be used by judges on these occasions. For example, as was the case in earlier studies of papers featuring more extended tasks (e.g., Greatorex et al., 2008, pp. 4–5), there is some evidence that performance on higher-tariff questions is attended to more, particularly as tiebreakers if the two candidates in a pair are close in quality. There is a sense here of these judges identifying a hierarchy of skills. In other words, if two candidates were relatively evenly matched in performance on most elements, they would be separated by their performance on the skills tested more in these higher-tariff questions, such as problem-solving, say. This may be a good thing, as long as it is done relatively consistently by judges, but if the higher-tariff questions are not testing the same skills or knowledge as the paper as a whole, the issue of certain parts of the paper playing an outsize role in judgements is a live one.

Indeed, if it is the case, as it seems to be in these studies, that CJ judges attend more to certain questions (such as those worth more marks, or those more related to problem-solving than recall, for example) than others, what does this mean for validity? The hypothetical situation where a script which had overall received fewer marks but was judged superior due to the judge preferring its writer's answers to problem-solving questions, for example, raises significant questions about the acceptability of comparative judgement-informed awarding processes in consistency terms. This situation is likely to be mitigated by the simplified pairs approach, which collates many judgements and regresses them against the scripts' mark difference, but this mitigation (which reduces the impact of any individual judge's inconsistency from the approach of others) may not be recognised by judges or other stakeholders. Furthermore, it might be seen as a good thing that judges concentrate on certain, better-discriminating, questions, if these can be seen as identifying the superior mathematician or scientist, say, more efficiently. However, there is certainly a potential tension here; ultimately, what *should* we be asking judges to decide their judgements on?

This issue is not unique to CJ in subjects relying on highly structured papers, but may be more pertinent in them. This is because papers using extended response tasks are likely to test the required skills in most, if not all, of their tasks, whereas highly structured papers may have one set of sections or items focusing on each required skill. In a marking schema, the sum of individual judgements of candidates' performance on these skills thus creates an overall mark which reflects their performance appropriately across all skills tested on the paper, but with holistic comparative judgements creating this overall judgement appropriately may be more of a challenge. What is important – both in marking and CJ – is that there is clarity as to what kind of skills are being tested when and why, and if there is meant to be a hierarchy of skills.

Re-marking

The issue of whether judges were simply re-marking the papers in front of them in accordance with their original mark schemes, and then selecting as superior the one they awarded the most marks to (in contrast to, as was intended, making true quick holistic judgements) is an important one for the validity of CJ.

Evidence from these studies suggests that, at least for some judges, reverting to re-marking was difficult to avoid. All the judges chosen for these studies were experienced markers of the relevant qualifications, and as such had been trained in performing the precise item-by-item determinations of right or wrong that are critical to how marking works in these contexts. The psychological transition to making CJ judgements is substantial. This is for two reasons. Firstly, a quick assessment being made of the *overall* quality of a paper, in a holistic fashion, is very different from the precise, standardised methods of marking. Secondly, an individual judge's decision matters less in CJ, in that CJ methods bring together the judgements of many. This situation (where judges do not need to act as though their judgement alone has to be right all the time) may be difficult for judges to adjust to. It may have been the case that this latter point was not understood well by judges, who were used, as markers, to their marking being decisive in a student's outcomes, and therefore expected to put a lot of effort into getting it right every single time.

This highlights the importance for the future, if CJ is to be rolled out in wider settings, and especially in STEM subjects and highly structured papers, of getting judge training right. CJ relies significantly on judges making their judgements *in the way we want them to*, but without necessarily telling them how. Judges with experience of marking need to be aided to make the transition to the CJ mindset – perhaps with training materials, testimonials from judges who have used CJ about how it works, and evidence of its appropriateness, as well as the opportunity to try the method and receive feedback. It should not be assumed that this is a trivial issue, as working under a CJ mindset may be seen by judges as a challenge to their professionalism as markers. CJ thus risks not being viewed as a desirable task, and then not getting the necessary examiner buy-in. However, evidence from these studies suggests that judges' ease with the process increased throughout the exercises – i.e., as they gained experience and knowledge about what they were doing – implying that this transition is possible to achieve.

CONCLUSION

So, can comparative judgement accurately estimate the true difference in difficulty between two exam papers comprising many highly structured tasks (RQ1)? The mathematics study reported on here suggests that the estimated difference derived from simplified pairs could closely approximate the empirical equating difference. However, in the science exercise, the CJ outcome did not closely align with the empirical difference in difficulty between the two papers. It is difficult to explain this discrepancy, though reasons may include the differences in the content of the exams or the specific judges. Nonetheless, based on the science exercise, we now know for certain that comparative judgement need not lead to an accurate impression of the relative difficulty of different exams. More research is needed to ascertain the particular conditions (if any) under which we can be confident that CJ can accurately estimate the true difference in difficulty between two exams of highly structured tasks.

We have also addressed the question of how judges make comparative judgements of students' work from exam papers comprising many highly structured tasks (RQ2). The processes that judges used to make decisions when judging papers made up of highly structured tasks were varied – with the same judge likely to use different processes throughout their work. However, on the whole, judges were generally consistent enough in their processes.

One strategy used by some judges working on highly structured papers was to make decisions based on a subset of the exam paper. The validity of CJ depends on judgements that are holistic because judgements made on a subset of the questions in an exam may omit some target constructs which, consequently, means that scripts may be being judged (for the purpose of assigning grades) against different criteria to those they are being marked against. This may not be acceptable as it is then unclear exactly what skills students are being assessed on. Moreover, those skills embodied in the mark schemes may be subtly different. Another strategy reported by judges was to re-mark the papers and then compare scripts based on a totting up of scores on the items in the paper. However, re-marking within a CJ exercise negates the benefit of speed. It also means that judges are not necessarily accounting for the differences in difficulty between 1 year's paper and the next. In all these cases, a greater understanding of what judges are doing when they judge comparatively brings to the fore questions of assessment judgement validity that generally remain implicit in the marking and non-comparative judgement contexts.

The strategies used in exam marking processes are well understood (e.g., Suto et al., 2008; Crisp, 2010a). This paper adds to our understanding of processes used by CJ judges when making decisions about highly structured papers. However, this area is still not as well theorised as that of decision-making in marking. More research to further this understanding and to build knowledge of the impact of judging decisions and processes on CJ outcomes would be welcome. Further research is also required into what is meant by a holistic decision, and how to manage the cognitive load that arises when judging student work which contains many short answer questions, so that exam boards can provide fuller guidance to judges about how they should make decisions in CJ tasks and what information in the papers they should be concentrating on.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TB conceptualized the simplified pairs method and designed the maths study. TL and SH organized the maths study. TG ran the science study on the same design. TL, TG, and SH wrote the first manuscript draft, with TG writing the section on study findings, TL the section on judge survey findings and the discussion, and SH the introduction and conclusion sections. All authors contributed to manuscript revision and read and approved the submitted version.

ACKNOWLEDGMENTS

We would like to acknowledge the contributions of colleagues in OCR, including Natalie Gawthrop, Charlotte Gow, and Stephen Furness, for supporting the operationalisation of the CJ work, and Terry Child for developing the CJ software tool.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2022.803040/full#supplementary-material

REFERENCES

Albano, A. D. (2016). equate: an R package for observed-score linking and equating. *J. Statistical Software* 74, 1–36. doi: 10.18637/jss.v074.i08

Baird, J.-A. (2007). "Alternative conceptions of comparability," in *Techniques for Monitoring the Comparability of Examination Standards*, eds P. E. Newton, J.-A. Baird, H. Goldstein, H. Patrick, and P. Tymms (London: Qualifications and Curriculum Authority). doi: 10.2307/j.ctv15r5769.4

Benton, T. (2021). Comparative judgement for linking two existing scales. *Front. Educ.* 6:775203. doi: 10.3389/feduc.2021.775203

Benton, T., Cunningham, E., Hughes, S., and Leech, T. (2020). *Comparing the Simplified Pairs Method of Standard Maintaining to Statistical Equating*. Cambridge: Cambridge Assessment. Cambridge Assessment Research Report.

Benton, T., Gill, T., Hughes, S., and Leech, T. (2022). A Summary of OCR's Pilots of the use of comparative judgement in setting grade boundaries. *Res. Matters Cambridge Assess. Publication* 33, 10–30.

Black, B. (2008). *Using an Adapted Rank-ordering Method to Investigate January versus June Awarding Standards*. Cambridge: Cambridge Assessment. Cambridge Assessment Research Report.

Bradley, R. A., and Terry, M. E. (1952). Rank analysis of incomplete block designs: i. the method of paired comparisons. *Biometrica* 39, 324–345. doi: 10.2307/2334029

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgement. *J. Appl. Meas.* 6, 202–223.

Bramley, T. (2012). The effect of manipulating features of examinees' scripts on their perceived quality. *Res. Matters: Cambridge Assess. Publication* 13, 18–26.

Chambers, L., and Cunningham, E. (2022). Exploring the validity of comparative judgement - do judges attend to construct-irrelevant features?. *Front. Educ.* 6: 802392.

Crisp, V. (2008). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge J. Educ.* 38, 247–264. doi: 10.1080/03057640802063486

Crisp, V. (2010a). Towards a model of the judgement processes involved in examination marking. *Oxford Rev. Educ.* 36, 1–21. doi: 10.1080/03054980903454181

Crisp, V. (2010b). Judging the grade: exploring the judgement processes involved in examination grading decisions. *Eval. Res. Educ.* 23, 19–35. doi: 10.1080/09500790903572925

Curcin, M., Howard, E., Sully, K., and Black, B. (2019). *Improving Awarding: 2018/2019 Pilots*. Ofqual report Ofqual/19/6575. Coventry: Ofqual

Gill, T., and Bramley, T. (2013). How accurate are examiners' holistic judgements of script quality? *Assess. Educ: Principles Policy Practice* 20, 308–324. doi: 10.1080/0969594x.2013.779229

Gill, T., Bramley, T., and Black, B. (2007). "An investigation of standard maintaining in GCSE English using a rank ordering method," in *Paper Presented at the Annual Conference of the British Educational Research Association*, (London).

Good, F. J., and Cresswell, M. J. (1988). Grade awarding judgements in differentiated examinations. *Br. Educ. Res. J.* 14, 263–281. doi: 10.1080/0141192880140304

Greatorex, J. (2007). "Contemporary GCSE and a-level awarding: a psychological perspective on the decision-making process used to judge the quality of candidates' work," in *Paper Presented at the Annual Conference of the British Educational Research Association*, (London).

Greatorex, J., Novakovic, N., and Suto, I. (2008). "What attracts judges' attention? a comparison of three grading methods," in *Paper Presented at the Annual Conference of the International Association for Educational Assessment*, (Cambridge).

Humphry, S., and McGrane, J. (2015). Equating a large-scale writing assessment using pairwise comparisons of performances. *Australian Educ. Research.* 42, 443–460. doi: 10.1007/s13384-014-0168-6

Jones, I., Swan, M., and Pollitt, A. (2015). Assessing mathematical problem-solving using comparative judgement. *Int. J. Sci. Math Educ.* 13, 151–177. doi: 10.1007/s10763-013-9497-6

Leech, T., and Chambers, L. (2022). How do judges in Comparative Judgement exercises make their judgements? *Res. Matters Cambridge Univ. Press Assess. Pub.* 33, 31–47.

Ofqual (2017). *Comparable Outcomes and New A Levels*. Available online at: https://ofqual.blog.gov.uk/2017/03/10/comparable-outcomes-and-new-a-levels/ (accessed 6 October 2021).

Suto, I., and Greatorex, J. (2008). What goes through an examiner's mind? using verbal protocols to gain insights into the GCSE marking process. *Br. Educ. Res. J.* 34, 213–122. doi: 10.1080/01411920701492050

Suto, I., Crisp, V., and Greatorex, J. (2008). Investigating the judgemental marking process: an overview of our recent research. *Res. Matters: Cambridge Assess. Publication* 5, 6–9.

Verhavert, S., Bouwer, R., Donche, V., and De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement, assessment in education: principles. *Policy Practice* 26, 541–562. doi: 10.1080/0969594x.2019.1602027

Wright, B. D., and Masters, G. N. (1990). Computation of OUTFIT and INFIT statistics. *Rasch Measurement Trans.* 3, 84–85.

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Education

ORIGINAL RESEARCH
published: 04 May 2022
doi: 10.3389/feduc.2022.784261

Assessing the Quality of Argumentative Texts: Examining the General Agreement Between Different Rating Procedures and Exploring Inferences of (Dis)agreement Cases

Yana Landrieu, Fien De Smedt, Hilde Van Keer and Bram De Wever*

Department of Educational Studies, Ghent University, Ghent, Belgium

OPEN ACCESS

Edited by:
Marije Lesterhuis,
Spaarne Gasthuis, Netherlands

Reviewed by:
Stefan Daniel Keller,
University of Applied Sciences and Arts Northwestern Switzerland, Switzerland
Gustaf Bernhard Uno Skar,
Norwegian University of Science and Technology, Norway

****Correspondence:***
Yana Landrieu
Yana.Landrieu@UGent.be

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *27 September 2021*
Accepted: *29 March 2022*
Published: *04 May 2022*

Citation:
Landrieu Y, De Smedt F, Van Keer H and De Wever B (2022) Assessing the Quality of Argumentative Texts: Examining the General Agreement Between Different Rating Procedures and Exploring Inferences of (Dis)agreement Cases. Front. Educ. 7:784261. doi: 10.3389/feduc.2022.784261

Assessing argumentative writing skills is not a straightforward task, as multiple elements need to be considered. In function of providing feedback to students and keeping track of their progress, evaluating argumentative texts in a suitable, valid and efficient way is important. In this state-of-the-art exploratory study, 130 argumentative texts written by eleventh graders were assessed by means of three different rating procedures (i.e., absolute holistic rating, comparative holistic rating, and absolute analytic rating). The main aim of this study is twofold. First, we aim to examine the correlations between the three rating procedures and to study the extent to which these procedures differ in assigning scores. In doing so, the more innovative approach of pairwise comparisons is compared to more established assessment methods of absolute holistic and analytic rating. Second, we aim to identify key characteristics that determine the quality of an argumentative text, independent of the rating procedure used. Furthermore, key elements of mid-range, weak and strong argumentative texts were studied in detail. The results reveal low to moderate agreement between the different procedures, indicating that all procedures are suitable to assess the quality of an argumentative text; each procedure, however, has its own qualities and applicability.

Keywords: argumentative writing, rating procedures, holistic rating, analytic rating, pairwise comparisons

INTRODUCTION

Effective writing skills are considered imperative in our twenty-first century society, as they are highly valued in private, educational, and professional contexts (Graham and Perin, 2007). This is especially true for argumentative writing skills. Argumentative writing skills are considered important as they help to clarify our thoughts and make us reflect on the thoughts of others (by integrating different points of view) and stimulate critical thinking and problem-solving competences (Varghese and Abraham, 1998; Nussbaum and Schraw, 2007; Granado-Peinado et al., 2019). However, the majority of students experience difficulties developing effective writing skills in general, and more particularly in the genre of argumentative writing (NCES, 2012). The argumentative writing proficiency of students appears to be highly substandard (Graham and Perin, 2007; NCES, 2012; Ferretti and Lewis, 2013; Song and Ferretti, 2013; Traga Philippakos and MacArthur, 2019). Ferretti and Lewis (2013), for example, found that students' argumentative texts rarely acknowledge opposing positions, rarely consider the merits of different views, and almost never include rebuttals of alternative perspectives.

Determining the quality of an argumentative text is not a straightforward task as different elements need to be considered. Nevertheless, with regard to providing feedback to students, keeping track of their progress, and helping them to write better texts, it is important to be able to evaluate argumentative texts in a suitable, valid, and efficient way. By taking a closer look at the texts in our sample, we have gained insights regarding features of stronger and weaker argumentative texts, which will be shared in this study. In what follows, we firstly present three rating procedures that are central in this study: (1) Absolute holistic rating, (2) comparative holistic rating, and (3) absolute analytic rating. As comparative holistic rating is an innovative and upcoming assessment procedure in writing research, we will compare this rating procedure to more established methods such as absolute holistic rating and absolute analytic rating. Next, we briefly review the literature on the assessment of argumentative texts. More specifically, we discuss the need to assess (1) the quality of argumentation, (2) the quality of content, and (3) the inclusion of general text characteristics to determine the overall quality of an argumentative text. The main aim of this exploratory study is to compare (a) different rating procedures that can be used when assessing argumentative texts, and (b) to identify text features of weak and strong argumentative texts. This study is innovative as this is the first study comparing the three rating procedures, especially given that pairwise comparisons are not yet as widespread and established as holistic and analytic rating. Secondly, we closely examine the specific features of a weak or strong argumentative text. Which features make a text weak or strong? Combining these two insights can be informative for assessment practices and give more insight into the key aspects of an argumentative text, regardless of the rating procedure used.

THEORETICAL BACKGROUND

Rating Procedures to Assess Text Quality

It is essential to assess the quality of argumentative texts in a suitable and valid way. Selecting a rating procedure is, however, not easily decided. In determining the most suitable procedure, a number of factors, such as the available time, the aim of the assessment, and the amount of raters and texts, should be taken into account. A review of previous research shows that many different procedures are used to assess written texts. These procedures differ in the degrees of rating freedom. Following Coertjens et al. (2017), rating procedures can be classified in two dimensions: Holistic vs. analytic on the one hand and absolute vs. comparative on the other hand (see also Harsch and Martin, 2013; Bouwer and Koster, 2016; Coertjens et al., 2017). In holistic rating, texts are rated as a whole, whereas in analytic rating, text quality is measured by scoring multiple features of a text. In absolute ratings, every text is scored by a description or a criteria list, whereas in comparative ratings, texts are compared to each other to assess the text quality. In this study, we focus on three rating procedures: (1) Absolute holistic rating, (2) comparative holistic rating, and (3) absolute analytic rating.

Absolute Holistic Rating

Within absolute holistic rating, there are differences regarding the extent to which a rater has access to specific rating criteria. For instance, a holistic rubric provides the rater with predefined rating criteria. In this way, raters using such rubrics still provide a holistic assessment based on their overall impression of a text but they are supported by the holistic explanations provided with each text score (Penny et al., 2000; Yune et al., 2018).

Another way to holistically assess a text is general impression marking. Following this procedure, texts are rated as a whole by assigning a score based on a total impression (Charney, 1984). Raters receive a general description regarding the assignment and the competences that are being pursued while writing. However, raters do not receive explicit rating criteria to assign a particular score. Each rater has (unconsciously) an internal standard on how to evaluate a text, *inter alia*, based on earlier rating experiences. An advantage reported in the literature is that this procedure does not require a lot of time and effort, as scores are rather quickly assigned without explicit rating criteria (Charney, 1984).

There are two drawbacks linked to general impression marking: rater variance and the lack of detailed feedback on students' performance (Carr, 2000; Weigle, 2002; Lee et al., 2009). Regarding rater variance, not every rater uses the full scale to assign scores. For instance, raters can vary in terms of rigor by systematically assigning either higher or lower scores to texts. Additionally, raters can also have different rating criteria in mind or can perceive some elements as more important than other elements, even though they are asked to rate holistically (Weigle, 2002; Lee et al., 2009; Bouwer and Koster, 2016). Another explanation for varying scores is the *halo effect*, as described by Thorndike (1920). "Ratings *are* apparently affected by a marked tendency to think of the person in general as rather good or rather inferior and to color the judgments of the qualities by this general feeling" (p. 25). The quality of general impression marking may also depend heavily on the experience of the assessors. Rater training and experience could increase the reliability between raters, but this is not automatically the case (Myers, 1980; Charney, 1984; Huot, 1993; Rezaei and Lovorn, 2010; Coertjens et al., 2017). To reduce rater variance, support (i.e., rater training, or support in using the whole scale) for holistic raters is thus essential. By doing so, the reliability of the ratings can be increased (Bouwer and Koster, 2016). However, when raters are supported with criteria, we no longer apply general impression marking, as this is a rating procedure that works without rating criteria. As to the second drawback, general impression marking does not provide insight into students' weaknesses and strengths in (argumentative) writing in detail. In this respect, a general score is assigned to a text without providing any detail or information on how and why this particular score was assigned.

Nevertheless, teachers can provide additional feedback so the student has insight into the strengths and weaknesses of the text.

Whenever absolute holistic rating is mentioned in this study, we are referring to general impression marking. We chose to implement absolute holistic rating in this way, as many teachers in schools still use this approach when evaluating argumentative texts and we were also able to observe this in Flanders.

Comparative Holistic Rating (Pairwise Comparisons)

In comparative holistic rating, texts are holistically compared to each other to assess the text quality. A well-known comparative holistic rating approach is pairwise comparison. The holistic character implies that raters are free to define how to assess the texts without any predetermined criteria (van Daal et al., 2016). The comparative character implies that each rater compares two texts and selects the best one. This is applied multiple times and creates a binary decision matrix of the worst and the best text in each comparison (Coertjens et al., 2017). This results in a reliable ranking order of texts ranging from the worst rated text to the best rated text. Texts are constantly compared to each other, and each text is evaluated multiple times, by multiple raters. This procedure is based on Thurstone's law of comparative judgment (1927) which explains how objects (e.g., written argumentative texts) can be scaled from lowest to highest text quality by pairwise comparisons (Pollitt, 2012). Following Thurstone (1927), raters are more competent in comparing two different texts to each other than to rate one text as a whole (Thurstone, 1927; Gill and Bramley, 2013; McMahon and Jones, 2015). Multiple raters compare two different texts and select the best one, according to their opinion. By using the Bradley-Terry model, a scale from worst to best text can be generated (McMahon and Jones, 2015; Coertjens et al., 2017). By using this scale, teachers or writing researchers can easily assign a score to a text. This method originated in psychophysical research but has become applicable for educational assessment purposes as well (Pollitt, 2012; McMahon and Jones, 2015).

This procedure is easy to implement, as raters simply have to decide which of the two presented texts is the best one (McMahon and Jones, 2015). By doing so, pairwise comparisons eliminate differences in the severity of raters (van Rijt et al., 2021). Another advantage is that there is no need for an extensive training procedure for raters. However, deciding which text is better can be difficult in some instances (e.g., a text with high-quality content, but with poor argumentative structure or two texts of a similar level). Therefore, raters need a clear understanding of the writing assignment goals to assess which text is the best one. Pairwise comparisons are not easily applicable in regular teaching activities, as multiple raters are required to achieve a reliable scale. Due to the fact that this procedure can be difficult to implement in a school context, this procedure is sometimes considered inefficient (Bramley et al., 1998; Verhavert et al., 2018).

Overall, the reliability of pairwise comparisons appears to be much higher compared to absolute holistic rating procedures (Thurstone, 1927; Pollitt, 2012; Gill and Bramley, 2013). The reliability of pairwise comparisons depends on the amount of comparisons: The more comparisons, the more reliable the ranking order (Bouwer et al., in review)[1]. Next to a high reliability, pairwise comparisons also provide valid scores (Pollitt, 2012; van Daal et al., 2016). Each text is evaluated by multiple raters and the final ranking order is a reflection of the multiple raters' expertise (Pollitt, 2012; van Daal et al., 2016). This implies that pairwise comparisons result in a reliable ranking order. Individual rater effects can be neglected, due to the large number of raters, which ensures that each text can be compared several times with another text (e.g., in this study each text is, on average, compared 16.6 times to another text).

Pairwise comparisons are not the only way to assess texts in a comparative, holistic way. Benchmark rating could also be a way of comparatively and holistically assessing an argumentative text, by providing raters benchmarks that each represent a certain text quality. For more information on this comparative holistic rating procedure, we refer the reader to Bouwer et al. (see text footnote 1). As we opted to use pairwise comparisons in this study, benchmark rating will not be further explored.

Absolute Analytic Rating

Analytic rating procedures are more detailed than holistic procedures, as text quality is measured by scoring multiple features of a text (i.e., sub scores for specific text features or facets that a rater has to keep in mind) which can be added up (Harsch and Martin, 2013; Coertjens et al., 2017). There are several advantages linked to this procedure. First, by using an analytic rating procedure, weaknesses and/or strengths in a text can be distinguished, leading to more information for teachers or researchers. This can lead to more precise feedback which can improve the learning process of the student (Lee et al., 2009). Second, earlier research (Vögelin et al., 2019) showed that lexical features can have an influence on how text quality is rated; however, the chance that one specific weakness in a text (e.g., grammar) is decisive in the overall assessment is smaller for analytic procedures than it is with holistic rating (Barkaoui, 2011). Third, by defining rating criteria in advance, more equal and reliable scores between raters can be obtained. Previous research on reliability of analytic rating is, however, still very inconsistent. Earlier research of Follman et al. (1967) and Charney (1984) claims that absolute analytic rating does lead to good or increased reliability compared to absolute holistic rating, whereas research by Goulden (1994) and Barkaoui (2011) claims that analytic rating leads to decreased reliability. In addition, training the raters could increase reliability and validity, but this does not automatically lead to reliable and valid scores (Rezaei and Lovorn, 2010; Harsch and Martin, 2013). Therefore, Harsch and Martin (2013) suggest combining holistic and analytic rating procedures to achieve more reliable and valid results.

In addition to the enumerated benefits, previous research also reported several drawbacks related to analytic rating. More specifically, analytic rating can be time consuming, as each text feature is separately scored (Hunter, 1996). In this respect, it is questionable whether the sum of the parts is a representative score of a text (Huot, 1990). When writing researchers or teachers want to achieve a reliable score, multiple raters can be used. This,

[1] Bouwer, R., Lesterhuis, M., De Smedt, F., Van Keer, H., and De Maeyer, S. (in review). *Comparative Approaches to the Assessment of Writing: Reliability and Validity of Benchmark Rating and Comparative Judgement*.

however, makes it more difficult to apply in practice (Lee et al., 2009). During the assessment, raters may not be able to call upon their expertise (and do not have ownership of the total score), as they are tied to rating the predetermined criteria (in contrast to absolute and comparative holistic rating). Additionally, analytic rating does not automatically create rich data, as all elements are simply added up (Hunter, 1996). In other words: Analytic rating does not look at the whole picture, as opposed to absolute and comparative holistic rating. Time must be invested in setting up an analytical rating procedure.

A frequently used analytic rating procedure is the use of rubrics. By using an analytic rubric, written texts are rated on multiple aspects and sub scores are allocated considering specific text features or facets that a rater has to keep in mind (Weigle, 2002; Barkaoui, 2011; Harsch and Martin, 2013). By adding up the sub scores, an overall score can be assigned. The goal of using the rubric-criteria is to enlarge the agreement between different assessors and thus reduce rater variability. In an analytic rubric, the text features are predetermined, but the weight of these text features is not always determined in advance. Following Sasaki and Hirose (1999) and Coertjens et al. (2017), raters can independently decide which weight they give to the text features. This implies that a text feature to which the rater attaches great importance can be more decisive than another text feature. Other authors, like Stapleton and Wu (2015), describe the weight of the separate text features in a rubric as fixed. This implies that the rater cannot decide the weight of each text feature and this makes analytic rating less free than holistic rating.

Determining the Overall Quality of an Argumentative Text

Quality of Argumentation

Toulmin (1958) argued that an argumentative text is composed of (a) a claim, (b) data, (c) warrants, (d) backings, (e) qualifiers and (f) rebuttals. The claim is the thesis of the author, whereas data is the foundation for the claim. A warrant is the relation between the data and the claim. Backings are additional evidence that support the claim. A qualifier adds credibility to the argument, whereas rebuttals are circumstances under which a claim is not valid (Toulmin, 1958). The original Toulmin-model has been modified in contemporary literature into a more understandable and practical model (Nussbaum and Kardash, 2005; Nussbaum and Schraw, 2007; Qin and Karabacak, 2010; Stapleton and Wu, 2015). Alongside the work of Toulmin (1958) and Stapleton and Wu (2015) stated that a strong argumentative text is composed of two important elements. First, an argumentative text must be constructed taking into account all elements contributing to a good *quality of argumentation.* Second, attention must be paid to the *quality of the content* of the text. According to Clark and Sampson (2007) and Stapleton and Wu (2015), many studies prioritize the importance of the quality of argumentation over the quality of content. As Clark and Sampson (2007) mention, the majority of research on argumentative writing skills focuses explicitly on the Toulmin-structure, without paying attention to the content of the argumentative text leading to an incomplete picture of the quality of the text (Simon, 2008). In line with Stapleton and Wu (2015), we therefore argue that it is not self-evident, but important to take both the quality of argumentation, the quality of the content and the general text characteristics into account when evaluating an argumentative text, as all three elements are connected and cannot be completely separated.

Quality of Content

In addition to the quality of argumentation, previous studies also examined the quality of the content. In this respect, three criteria are distinguished in the literature: overall persuasiveness, factual accuracy, and information originating from source texts. First, as the main goal of argumentation is to convince or persuade an audience of a certain point of view, a high-quality argumentative text should have a good overall persuasiveness (De La Paz and Felton, 2010). Strong persuasive arguments require deep reasoning from students, as they need to come up with good reasons to support the claim (Marttunen et al., 2005). Second, an argumentative text should be factually accurate (De La Paz and Felton, 2010). Third, the author should integrate information originating from multiple, reliable source texts into one's argumentative text (De La Paz and Felton, 2010; Cuevas et al., 2016). This implies that the author needs to consider the multiple points of view that are present in the source texts (Wolfe and Britt, 2008). Writers must have the capacity to draw upon evidence to support their point of view (Kibler and Hardigree, 2017). It is allowed for writers to express their own opinions, but it is recommended that they support these opinions with objective sources.

General Text Characteristics

As well as considering the quality of argumentation and the quality of content, various general text characteristics also appear to be key in determining the overall quality of an argumentative text. More particularly, including an *introduction* and/or *conclusion* in a text can be helpful for the reader. A good introduction draws the reader's attention and reveals the main topic of the text to the reader, and by reading the conclusion, readers can quickly find out the point of view of the author (Syed et al., 2021).

In addition, as Barkaoui (2010) and Wolfe et al. (2016) mention, *text length* significantly influences text quality. Longer texts contain more information and details and are therefore often associated with a higher text quality. However, including unnecessary and *irrelevant information* in texts can hinder the flow and readability of a text. Finally, *bad writing mechanics* seem to negatively affect text quality (Figueredo and Varnhagen, 2005; Rezaei and Lovorn, 2010; Jansen et al., 2021). However, this list is not exhaustive. There are many other elements (e.g., structure, logical line of reasoning, etc.) that determine text quality, but these are out of the scope of this study.

The Present Study

A variety of assessment methods exists, but the literature generally distinguishes between holistic and analytic rating procedures, as discussed in the theoretical background. There appears to be a misconception that the use of analytic rating automatically leads to a reliable score. As the results in the educational field seem to be inconsistent and reveal mixed results on reliability and validity (e.g., Charney, 1984; Barkaoui, 2011), more research is needed. Harsch and Martin (2013) reveal that both holistic and analytic rating procedures have their strengths and weaknesses, depending on the purpose for which they are used (see also Barkaoui, 2011). More recently, another distinction in rating procedures has been identified in the literature: absolute and comparative rating procedures (Coertjens et al., 2017). In this respect, a comparative approach by means of pairwise comparisons has been introduced to effectively and efficiently assess students' writing performance (Coertjens et al., 2017). Pairwise comparisons are proven to be a valid and reliable rating procedure and therefore seem to be a promising alternative for absolute holistic and absolute analytic rating procedures (van Daal et al., 2016; Coertjens et al., 2017). To date, there is no research yet that focuses on comparing these three rating procedures. Therefore, this study will tackle this issue. The main aim of this study is twofold. The first aim of this study is to examine the correlations between the three rating procedures and to study the extent to which these procedures differ in assigning scores. In doing so, the innovative approach of pairwise comparisons is compared to more established assessment methods of absolute holistic and analytic rating.

In this study, we choose to use pairwise comparisons as a starting point for describing results which we then use to make connections to the other rating procedures. There are three reasons for this approach. First, pairwise comparisons use multiple raters, leading not only to a high level of reliability, but also to a broadly based consensus. Research of Verhavert et al. (2018) showed that the Scale Separation Reliability (SSR) can be interpreted as an inter-rater correlation. Second, whereas holistic and analytic rating are more established and more often used in practice, pairwise comparisons are already commonly used in educational research and are considered promising methods to assess writing performance (Coertjens et al., 2017; Verhavert et al., 2018). The rating procedure is easy to implement for researchers, as specific software exists and raters do not need a lot of training, and it provides opportunities to achieve high inter-rater reliability. However, it also requires a lot of different raters, so this rating procedure is less suitable to use in daily practice. Third, in this study, the use of pairwise comparisons is a procedure that takes into account both quality of content and quality of argumentation. These three arguments ensure that this rating procedure is an optimal procedure to start from and to compare to the other two rating procedures.

The second aim of this study is to identify key characteristics that determine the quality of an argumentative text, independent of the rating procedure used. Regarding the second aim, in addition to making an informed choice regarding the assessment procedure, the evaluator must also have an understanding of the essential criteria of an argumentative text. Based on previous research by Stapleton and Wu (2015), the overall quality of an argumentative text is determined by the quality of argumentation and the quality of content. In addition, several general text characteristics (e.g., the inclusion of an introduction and conclusion, text length, use of irrelevant information and writing mechanics) should be taken into account as they influence (argumentative) text quality as well. Therefore, we want to identify key characteristics that determine the quality of an argumentative text. In this respect, we particularly focus on examining the elements that seem to be associated with mid-range, weak and strong argumentative texts. Based on the twofold aim of the study, three main research questions are addressed in the present study.

RQ1a: How do absolute holistic rating, comparative holistic rating (pairwise comparisons) and absolute analytic rating correlate?

RQ1b: How often do we see deviations between these rating procedures and how strong are these deviations?

RQ2: Which elements characterize mid-range, weak and strong argumentative texts, independent of the rating procedure used?

MATERIALS AND METHODS

Participants

In total, 164 eleventh grade students participated in the study and wrote an argumentative text. Students were on average 17 years old, their age varying between 16 and 19 years. All students were enrolled in the academic track of secondary education. The majority of the students were native Dutch speakers (*n* = 156, 95.1%), 3.7% were bilingual (Dutch + another home language) (*n* = 6) and 1.2% had another home language (*n* = 2) (French). The majority of the participants were female (*n* = 123, 75%).

Data Collection Procedure

After signing an active informed consent (the parents/guardians received a passive informed consent), the students had to complete an argumentative writing test. Half of them (*n* = 79) completed a digital writing test on the conservation of zoos, and the other half completed a digital writing test on voting rights from the age of 16 (*n* = 85). Each student received two source texts on the respective topic and was instructed to write an argumentative text based on the source texts and based on their own opinion. This integrated writing task required the secondary school students to write an argumentative text (with the goal to persuade the reader) by using the informative source texts. They were free to choose their own point of view and (counter) arguments and rebuttals. The secondary school students were not allowed to copy-and-paste from the source texts, but they were asked to integrate the arguments from the informative source texts into their own argumentative texts (in their own words). They were free to add additional arguments or other information, not directly drawn from the source texts. They were allowed to use a digital draft sheet, but were not allowed to search for extra information on the internet. The source texts were similar in difficulty and length (i.e., on average 634 words). Furthermore, the students were instructed to clearly take a stand and defend one position. They had to write individually and had to complete the argumentative writing test within 45 min, without further guidance.

Due to the Covid-19 pandemic, the data collection was discontinued abruptly. Nevertheless, we were able to collect 164 texts in total, of which 157 texts were further included in the study (i.e., due to late submission, seven texts could not be assessed using the three rating procedures). Although the assignment explicitly stated to write an argumentative text, 27 texts did not take a position (e.g., pro or contra), did not have the goal to persuade, nor were any arguments integrated. Therefore, these texts were categorized as informative and eliminated from further analyses. 130 argumentative texts with an average length of 401 words (*SD* = 113, min = 166, max = 873) were included in the analyses. All texts were anonymized. Raters were unaware of the gender and language background of the authors of the texts.

Rater Training and Rater Procedures

Raters

In light of a research assignment on assessment, university students enrolled in the second year of educational sciences (*n* = 132) collected the data. Prior to the data collection, the definition and the goal of argumentative writing were explained to the university students, and they were introduced to the differences between rating procedures. Furthermore, they received a protocol outlining the data collection procedure, which they had to follow strictly. After collecting the data, these 132 university students also served as raters for the pairwise comparisons. The holistic and analytic rating procedure were executed by the researcher and a trained rater (*n* = 2) (see **Table 1**).

Instructions for Holistic and Analytic Rating

The argumentative texts were holistically and analytically rated by the first author and a trained language teacher who teaches Dutch in secondary education (see **Table 1**). According to Bacha (2001), training additional raters in how to assess texts is key. Therefore, the second rater received an instruction guide and followed an intensive training session given by the first author (3–4 h). During this session, the structure of an argumentative text was explained in detail. More particularly, the adapted model of Toulmin, as used by Nussbaum and Kardash (2005) and Stapleton and Wu (2015), was instructed and each element of the model was illustrated by means of specific examples. Furthermore, both holistic and analytic rating were explained in detail and the specific assessment procedures were discussed and practiced. During practice, ten texts (on both writing topics) were rated holistically and analytically and discussed with the first author.

For holistic rating, no specific instructions were given to the rater except for the instruction to assign a holistic score from 0 to 10 that best reflects the quality of this argumentative text. The goal was to intuitively map the quality of the text according to a general impression without predefined criteria, as Myers (1980) recommends.

For analytical rating, the raters used the framework developed by Stapleton and Wu (2015), the so-called “Analytic Scoring Rubric for Argumentative Writing” (ASRAW). In the ASRAW, quality of argumentation is determined by looking at six elements, based on the earlier research of Nussbaum and Kardash (2005) and Qin and Karabacak (2010). The elements are: (a) A claim, (b) claim data, (c) a counterclaim, (d) counterclaim data (e), rebuttals, and (f) rebuttal data. **Table 2** provides an overview of these elements, including a description for each element. Ideally, all elements are included in a logically structured argumentative text. So the more a text conforms to the (adapted) Toulmin-structure, the stronger and more persuasive it can be (Qin and Karabacak, 2010). However, when a text does not include all elements, the text is not automatically considered a weak text. Much also depends on the quality of the content and the general text characteristics (Stapleton and Wu, 2015). The order in which the elements appear is neither linear nor predetermined (e.g., a text does not have to start with a claim, the counterclaim and counterarguments can be placed before the actual claim).

The ASRAW uses different performance levels (for claim data, counterargument data and rebuttal data) and a dichotomous scale (for claims, counterargument claims and rebuttals). Each rating dimension is given a score, and although the weight of the elements is predetermined, not all elements are given the same weight (e.g., if a text mentions a claim, a score of 5 is given; if a text mentions a counterargument claim, a score of 10 is given). The specific weight attached to each element was decided by Stapleton and Wu (2015), the original developers of the framework. As data, counterarguments, and rebuttals require a higher level of critical thinking and argumentation skills, a higher weight is given to these elements. By adding up the scores, a total score is presented for the whole argumentative text. Scores ranged from 5 to 100. For more detailed information, we refer to Table 4 in the original work from Stapleton and Wu (2015). As mentioned in the literature overview, the ASRAW seems to prioritize quality of argumentation over quality of content. For instance, a text that does not provide any data (i.e., arguments that defend the point of view) is automatically assigned score “0” for that element, whereas the content of the text might be good. Without a solid argumentative structure, an argumentative text can never receive a high final score according to the ASRAW.

Instructions for Pairwise Comparisons

Argumentative texts were assessed through pairwise comparisons by 132 university students (see **Table 1**). The platform Comproved (Comproved.com) was used to make the comparisons. Pollitt (2012) argues that raters do not need much training when comparing texts to each other (see also Coertjens et al., 2017). Therefore, only a few instructions were given to the raters. The instructions were: “Hen judging which argumentative text is the best one, you can keep the following criteria in mind: (1) The author takes a reasoned position, (2) the author substantiates the position with relevant arguments, (3) the author uses information from sources or presents their own reasoning to support their position, and (4) the text is comprehensible (cf., coherent text structure, sentence structure and word choice).” Correct spelling, use of punctuation and capitalization were not taken into account in the assessment. Alongside these instructions, the raters were also instructed on the genre of an argumentative text by providing them with a definition of argumentative writing and explaining the goal of this genre (i.e., persuading). Given the holistic and comparative nature of this assessment, we did not provide further explicit instruction on the different elements of strong argumentative texts to avoid raters checking for each Toulmin-element in an analytic way.

TABLE 1 | Overview of the used rating procedures, the amount of raters, and the assessment methods.

Rating procedure	Amount of raters	Assessment method
Holistic rating	*n* = 2 (The researcher and a trained rater)	General impression marking, without predefined criteria
Analytic rating	*n* = 2 (The researcher and a trained rater)	By the use of the ASRAW (Stapleton and Wu, 2015)
Pairwise comparisons	*n* = 132 (132 second year educational sciences students)	By the use of the platform Comproved (Comproved.com)

During the rating process, raters were shown two texts each time and they had to select which one was the best argumentative text. Each text was rated multiple times, by multiple raters. More particularly, each student rated 20 pairs of texts independently at home and each text was compared on average 16.6 times to another text. The informative texts (i.e., texts missing a position and arguments) were then eliminated from the data and a ranking from the worst rated text to best rated text was calculated.

Procedures to Obtain Inter-Rater Reliability

Holistic and Analytical Rating

After the training, both the first author and the second rater assessed texts individually and independently. The assessment followed a two-stage process. During the first stage, all texts were rated holistically. During the second stage, texts were rated analytically but in a different order and with 1 week in between to avoid dependency between the two procedures.

For both holistic and analytic rating, the first author rated all argumentative texts (*n* = 130 texts) and the second rater double coded 24% (*n* = 31) of the texts. The Intraclass Correlation Coefficient (ICC) of the holistic and analytic ratings was examined based on the two-way mixed model, measuring consistency between raters. For analytic rating, the ICC of the total score of the ASRAW was 0.98, while the ICC for holistic rating was 0.48 (for more detailed per rating procedure information, see **Table 3**).

There are large discrepancies between the ICCs of the holistic and analytic rating procedure. The analytic rating procedure (the ASRAW) appears to be a reliable way to assign scores to argumentative texts. The units of analyses were indicated in advance, which made it easier and more transparent for the rater to assign subscores (as each unit represents an element of an argumentative text), which could partially explain the high ICC. In the holistic rating procedure (general impression marking), we observe a low ICC of 0.48, which is in line with our predictions, as this is an intuitive score, assigned without predefined criteria.

TABLE 2 | Overview of the elements of an argumentative text with a definition of each element, based on Stapleton and Wu (2015).

Elements of an argumentative text	Definition
Claim	An assertion or opinion to a specific topic
Claim data	Data that supports the actual claim
Counterclaim	The possible opposing views contrary to the own claim
Counterclaim data	Data that supports the counterclaim
Rebuttal	A claim that refutes the counterclaim, by responding to the counterclaim
Rebuttal data	Evidence to support the rebuttal

TABLE 3 | Reliability measures per rating procedure.

Collected texts	
Holistic rating	ICC = 0.48
Analytic rating	ICC of the total score of the ASRAW = 0.98 ICC of the individual elements of the ASRAW: • Claim: ICC = 1 • Claim data: ICC = 0.91 • Counterargument: ICC = 0.85 • Counterargument data: ICC = 0.98 • Rebuttal: ICC = 1 • Rebuttal data: ICC = 0.95
Pairwise comparisons	SSR = 0.83

Pairwise Comparisons

After all texts were rated in Comproved, a rank order was generated ranging from the lowest to the highest text quality. In this way, a logit score for each text was estimated. The higher the logit score, the better the text. Research by Verhavert et al. (2018) states that Separation Scale Reliability (SSR) is a good way to check the inter-rater reliability as it can estimate the level of agreement between the multiple raters. SSR is derived from Rasch modeling and is, according to Verhavert et al. (2018), typically used as a reliability measure. SSR is comparable to the ICC for multiple raters, both reflecting reliability of average scores across raters (Verhavert et al., 2018; see text footnote 1). An SSR of 0.80 and higher indicates a high inter-rater reliability. In this study we obtained an SSR of 0.83 (see **Table 3**).

Data Analysis

Preparatory Analyses

Given that the majority of our participants were female and native Dutch speakers, preparatory analyses were conducted to study the relationship between home language and text quality on the one hand, and the relationship between gender and text quality on the other hand. Based on ANOVA analyses, results showed no significant relationships between home language and text quality [pairwise comparisons: $F_{(1,\ 156)} = 0.09$, $p = 0.76$; holistic: $F_{(1,\ 162)} = 0.33$, $p = 0.57$; and analytic: $F_{(1,\ 162)} = 0.31$, $p = 0.58$]

TABLE 4 | Scoring "writing mechanics" of an argumentative text.

	Score 2	Score 1	Score 0
Writing mechanics	> 2 Spelling errors and >2 Syntax errors	1–2 spelling errors or/and 1–2 syntax errors	No spelling errors and No syntax errors

nor between gender and text quality [pairwise comparisons: $F_{(1,\ 156)} = 0.82$, $p = 0.37$; holistic: $F_{(1,\ 162)} = 0.31$, $p = 0.58$; and analytic: $F_{(1,\ 162)} = 0.46$, $p = 0.50$]. In addition, given that text length and writing mechanics are key predictors of text quality, both variables were taken into account in the analyses. For text length, the number of words were counted. For writing mechanics, the following scoring was applied (see **Table 4**). Both text length and writing mechanics were not double coded, as evaluating them was not ambiguous.

The relationships between text length and text quality (for each rating procedure) on the one hand, and writing mechanics and text quality (for each rating procedure) on the other hand, were all significant except for the relation between text length and analytical text quality. Variance explained by (1) text length, (2) writing mechanics, and (3) a combination of both was, respectively, 28.9, 5.5, and 37.5% for pairwise comparisons, 2.6, 3.1, and 6.4% for the analytic rating procedure, and 8.7, 5.54, and 15.8% for the holistic rating procedure.

Furthermore, results of the preparatory analyses showed that the explained variance of text length for pairwise comparisons (28.9%) was the highest and quite substantial. A possible explanation might be that pairwise comparisons are more prone and sensitive to text length, as longer texts were often rated as more qualitative and better texts. In educational research, text length has often been proven to have a significant relationship with text quality (Jarvis et al., 2003; Lee et al., 2009). However, when comparing texts to each other (like pairwise comparisons do), text length is an easy criterion to use. After all, this is the first visual indicator you see when you are presented with text A and text B. With the absolute rating procedures (absolute holistic and absolute analytical rating) you do not have this foundation of comparison. In addition, absolute analytic scoring (the ASRAW) may be less prone to this, due to the specific criteria that are not focusing on text length.

Main Analyses

Main Analyses in View of RQ1a + RQ1b

To study RQ1a and RQ1b, general analyses were conducted on all 130 argumentative texts. To analyze the results, correlations and attenuated correlations were calculated for RQ1a. Concerning the correlation between rating procedures, Bouwer and Koster (2016) stated that: “Since the rating procedures will not have a perfect reliability due to measurement error, correlations between scores from two rating procedures will suffer from attenuation.” (p. 43). Therefore, we conducted corrections on the correlations to deal with unreliability and to reflect the true correlations between rating procedures (Bouwer and Koster, 2016). More specifically, we divided the observed correlation coefficient by the product of the square roots of the two relevant reliability coefficients (Lord and Novick, 1968; Bouwer and Koster, 2016). For RQ1b, an alluvial plot was developed to visualize the results.

Main Analyses in View of RQ2

To investigate RQ2, a content analysis (on all texts, $n = 130$) and an in-depth analysis (on a subsample of texts, $n = 15$) were conducted. As to the content analysis, units of meaning were used to divide each text into multiple units of analysis. A unit of meaning can be a phrase, sentence or paragraph corresponding to one of the elements of an argumentative text (e.g., a rebuttal). The segmentation into units of analysis was executed by the first author. In total, 1,437 units of analysis were coded. Each unit of analysis is linked to one code, varying from 1 to 9. See **Table 5** for an overview of the codes assigned to each unit of analysis and an example of each code. **Table 5** is a representation of the code book that was developed. The code book provided detailed information concerning argumentative text characteristics and general text characteristics. To support the raters (the first author and a trained second rater), various examples and exceptions were also included in the code book.

In the content analysis, a second rater double-coded 24% ($n = 31$) of the collected, argumentative texts ($n = 130$). Within the 31 double-coded texts, 369 units of analysis were double-coded. Krippendorff’s alpha was calculated to estimate the inter-rater reliability (Krippendorff and Hayes, 2007). The results indicate that the inter-rater reliability was high (α = 0.93).

For the in-depth analysis, we selected a subsample of argumentative texts. By means of the preceding analyses of RQ1, several argumentative texts could be perceived as “mid-range,” “weak” or “strong,” independent of rating procedure used. For the purpose of this study, we define a mid-range text as a text in the middle 40–60% across rating procedures. A weak text is defined as a text in the lowest 20% of each rating procedure, and likewise, a strong text is a text that scores in the top 20% across rating procedures. All argumentative texts ($n = 130$) were ranked from highest to lowest for each rating procedure. We were then able to identify the top 20% and bottom 20% for each procedure. Next, it was examined which specific texts were always (regardless of rating procedure) in the top 20% (i.e., 7 texts) and bottom 20% (i.e., 7 texts). We applied the same process with the mid-range

TABLE 5 | Overview of the codes corresponding to each unit of meaning.

Code	Element	Example
Structure of argumentation		
1	Claim	*Zoos must be kept open.*
2	Claim data/argument	*Animals in zoos live longer and safer.*
3	Counterclaim	*Some people have the opinion that zoos should be closed.*
4	Counterclaim data	*As animals who are living in zoos are suffering from a lack of surface area.*
5	Rebuttal	*Living a longer and safer life is more important to me than having a lot of surface area.*
6	Rebuttal data	*By living longer and safer, almost extinct animal breeds have more opportunities to reproduce.*
General text characteristics		
7	Introduction	*The debate about whether zoos should close has been going on for some time. Several animal rights organizations have already taken action and protested. In this text, I will argue and defend my opinion on this conflict.*
8	Conclusion	*So from this I conclude that animals should actually be allowed to live in zoos.*
9	Irrelevant information	*I have already visited 5 zoos, situated in Antwerp, Brugelette, Mechelen, Vleteren and Ghent.*

texts. However, there was only one text that scored each time, across rating procedures, in the middle 40–60%. In this way, we arrived at the current selection of weak (*n* = 7), strong (*n* = 7) and mid-range (*n* = 1) texts.

This subsample of texts (*n* = 15) was subjected to an in-depth analysis in three different areas: (a) Structure of argumentation, (b) quality of content and (c) general textual characteristics. By means of this in-depth analysis, we try to uncover the characteristics of texts that have been scored as mid-range, weak and strong (see **Table 6**).

First, the structure of the mid-range, weak and strong argumentative texts was closely examined. Based on the content analysis, we checked which specific, argumentative elements were present in the mid-range, weak and strong argumentative texts. As mentioned in the see section "Data Analysis," an argumentative text ideally includes all argumentative elements. Earlier research also showed that often in weak argumentative texts, only a claim, and arguments supporting that claim, are provided implying that the author of the text is affected by tunnel vision and is ignoring the other point of view/counterclaim (Wolfe and Britt, 2008; Ferretti and Lewis, 2013). By means of the content analysis, we thus examined the argumentative structure of the fifteen texts in depth (e.g., how many of these texts consist only of a claim and claim data? If counterarguments are given, are they always refuted?). Second, the quality of the content was studied. In the theoretical background, we clarified that an argumentative text ideally has a strong persuasiveness, good factual accuracy and uses information originating from the source texts. All mid-range, weak and strong argumentative texts were analyzed on their quality of content by examining these three elements. See **Table 7** for more coding details. Third, to examine general textual information of this subsample of texts, the content analysis was used to determine whether the fifteen mid-range, weak and strong texts contain an introduction, conclusion, and/or irrelevant information. In addition, text length and writing mechanics (already taken into account in the preparatory analyses) were included in this in-depth analysis.

RESULTS

RQ1a: How Do the Three Rating Procedures Correlate?

Correlations between the three rating procedures are moderate to high, and are all significant at the 0.001 level (see **Table 8**). The correlations show that the different procedures are positively correlated, but are not fully aligned so they may focus on different text characteristics. Following Bouwer and Koster (2016), corrections on the correlations were conducted to deal with unreliability and to reflect the true correlations, as described in "Data Analysis" section.

TABLE 6 | In depth-analysis on mid-range, weak, and strong argumentative texts (*n* = 15).

Analyses on:	By means of:
Structure of argumentation	Content analysis
Quality of content	Analyses on persuasiveness, factual accuracy and use of information originating from source texts
General textual information	Content analysis and analyses on text length and writing mechanics

RQ1b: How Often Do We See Deviations Between Rating Procedures and How Strong Are These Deviations?

Knowing that correlations between the three rating procedures are moderate to high (and all significant at 0.001 level, it is interesting to inspect the descriptive statistics. In **Table 9** the descriptive statistics of the assigned scores using the three different rating procedures have been listed.

As the procedure of pairwise comparisons is our starting point from which connections are made to the other rating procedures (see section "The Present Study" of this study), a distinction was made between texts that were assigned a low score on pairwise comparisons (lowest 20%) but a high score (top 20%) on both of the other procedures (both analytic and holistic rating) and vice versa. First, not a single text was identified with a low score on pairwise comparisons, but a high score on the other rating procedures. Second, only one text was found rated as top 20% for pairwise comparison and bottom 20% for both holistic and analytic rating.

In an alluvial plot, the scores of the three rating procedures are compared to one another (see **Figure 1**). As can be observed, not all rating procedures arrive at the same ranking order. This indicates that each procedure has a specific focus. For instance, a text with a low holistic score does not necessarily have a low score on the other two rating procedures. Based on the inspection of the alluvial plot, there are some texts that are systematically ranked among the lowest or highest by all three of the rating procedures. However, there are also a large amount of texts that were evaluated rather differently by the three rating procedures indicating that in general the rankings fluctuate among the three rating procedures. This means that, though the correlations are in general positive and significant, the three rating procedures do not lead to exactly the same rankings.

RQ2: Which Elements Characterize Mid-Range, Weak, and Strong Texts?

Based on the in-depth analyses of the subsample, multiple elements characterizing argumentative texts were repeatedly identified among the mid-range, weak and strong texts. In the next section, the most common elements are described and analyzed.

Elements of a Mid-Range Argumentative Text

As we see many mid-range texts in each rating procedure, only one argumentative text was found which scored in the middle 40–60% each time, independent of the rating procedure being used. Mid-range argumentative texts do not have the highest scores but do not score particularly low either. In **Table 10**, the elements of the mid-range argumentative text are summarized. In general, the

TABLE 7 | Coding of quality of content.

Persuasiveness	Weak: The reasons the author puts forward are (a) not profound and (b) insufficient. Average: The reasons the author puts forward are (a) not profound or (b) insufficient. Strong: Deep reasoning from students. The author comes up with (a) profound and (b) sufficient reasons to support the claim.
Factual accuracy	Bad: Incorrect information was found at least twice in the text Average: Incorrect information was found once in the text Good: All information provided by the author was correct
Information originating from source texts	Never: No information in the author's text originated from the source texts Sometimes: Some information in the author's text originated from the source texts Always: All information in the author's text originated from the source texts

results reveal that the structure of this text was not great (i.e., no rebuttals were included, no rebuttal data were given), although the content and the general composition of the text was quite good. This explains why the text was, across all rating procedures, situated in the middle 20% (40–60%). As this is a single text, we will not go further into detail.

Elements Weak Argumentative Texts

Seven argumentative texts were found to score at the lowest 20% in the dataset according to all rating procedures. In **Table 11**, the elements of weak argumentative texts are summarized. If an element is observed in four or more out of the seven weak texts, we consider this a key element. The results reveal a weak argumentative text is characterized by: (a) The inclusion of only a claim and argument(s), (b) tunnel vision, (c) weak factual accuracy, (d) a lack of information from source texts, (e) weak persuasiveness, (f) the inclusion of irrelevant information, (g) short text length, and (h) weak writing mechanics.

TABLE 8 | Correlation coefficients between holistic rating, analytic rating and pairwise comparisons.

	Argumentative texts (*n* = 130)		
	Holistic	**Analytic**	**Pairwise comparisons**
Holistic	–	0.577**	0.483**
Analytic	0.913**	–	0.298**
Pairwise comparisons	0.818**	0.336**	–

***Correlation is significant at the 0.001 level. Above the diagonal are the original correlations, below the diagonal are the attenuated correlations.*

TABLE 9 | Descriptive statistics of the assigned scores using the rating procedures.

	HOLISTIC RATING (on a scale from 0 to 10)	ANALYTIC RATING (on a scale from 0 to 100)	PAIRWISE COMPARISONS (Logit scores)
M	5.77	50.19	0.07*
SD	1.44	19.36	1.135
Median	6	55	0.17
Minimum score	2	15	–3.46
Maximum score	9	85	2.47
Range	7	70	5.93

**In pairwise comparisons the mean of the logit scores is usually equal to zero. However, in this study, this score slightly deviates as the informative texts were left out of the ranking.*

Elements of Strong Argumentative Texts

Next to mid-range and weak texts, it was examined whether there were texts that are rated as the 20% strongest texts independent of rating procedure. Seven texts were found and, as can be seen in **Table 12**, several text features can be associated with a strong argumentative text. All these text features appeared in a minimum of four out of seven strong texts. More specifically, the results reveal that strong argumentative texts are characterized by (a) the use of a claim, arguments, a counterclaim, counterarguments, rebuttals, and rebuttal data, (b) all counterarguments are refuted by rebuttal(s), (c) the integration of information from source texts, (d) strong persuasiveness, (e) factual accuracy, (f) use of an introduction, (g) use of a conclusion, (h) high number of words, and (i) good writing mechanics.

DISCUSSION

The present study focused on providing insight into three different rating procedures by studying similarities (correlations) and deviations between scores assigned by each rating procedure. We argue that all three rating procedures are suitable for evaluating argumentative texts. However, when comparing the three procedures, we notice that in general, the rankings fluctuate among the three rating procedures. All three procedures can be seen as proxies for the quality of the argumentative texts, however, they have their own approach and focus. In addition, we found several elements of argumentative texts that seem to be associated with mid-range, weak or strong texts. In the discussion, these elements will be further explored. We aim to guide practitioners, researchers, and teachers in choosing a suitable rating procedure by verifying the purposes for which certain procedures work well. The discussion is structured according to the three research questions and, at the end, the findings are compiled and translated into practice. Limitations and suggestions for follow-up research are also discussed.

RQ1a: How Do the Three Rating Procedures Correlate?

Regarding the first research question (RQ1a), we found that the three rating procedures (i.e., absolute holistic rating, comparative

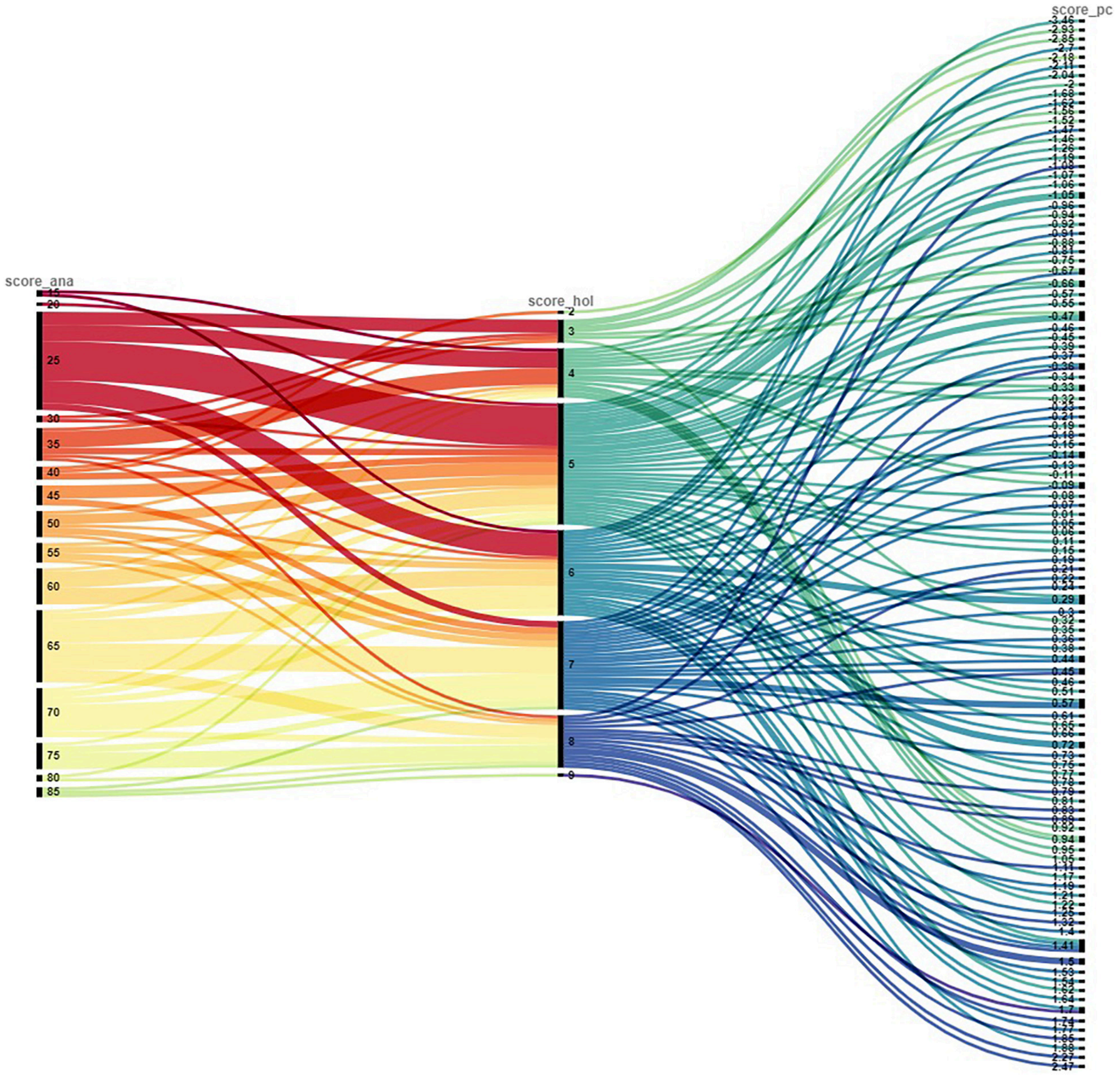


FIGURE 1 | Alluvial plot presenting the comparison of scores assigned through different rating procedures. Each colored line represents an argumentative text. The Figure shows a combination of the three procedures, with the analytic score on the left, the holistic score in the middle, and the pairwise comparisons score on the right. The scores per rating procedure are ordered from low (top of the alluvial plot) to high (bottom of the alluvial plot). Lines that are situated at the bottom of the figure represent a low score. Lines at the top of the figure represent a high score. The numbers on the vertical axes represent the scores attained by the rating procedure. Texts do not systematically score high or low on all three procedures, as becomes clear when looking at the figure.

holistic rating, and absolute analytic rating) correlate moderately to highly. Given that all procedures focus on assessing the quality of argumentative texts, this was in line with our expectations. However, the correlations are not fully aligned. Fully aligned correlations would indicate that rating procedures measure the underlying construct in exactly the same way (Messick, 1989). In this respect, the construct measured in this study is "argumentative writing skills."

This study revealed high attenuated correlations between absolute holistic and absolute analytic rating. When delving into the research literature, the findings on the correlations between both rating procedures are inconsistent. Studies by Freedman (1981) and Veal and Hudson (1983) show that holistic and analytic rating procedures correlate strongly. In contrast, studies of Hunter (1996) and Lee et al. (2009) indicate that holistic and analytic scores are not always and automatically strongly correlated. Keeping this contradiction in mind, in this study, we did not expect holistic and analytical rating to be this highly correlated because of the different focus of each procedure (i.e., holistic rating is based on the whole text whereas analytical rating focuses on specific argumentative text features). This high correlation could also be associated with the implementation of

TABLE 10 | Elements of the mid-range argumentative text.

	Elements mid-range argumentative texts
Quality of argumentation	Structure of the text: claim (1)—argumentation (3)—counterclaim (1)—counterargument (1) (no rebuttals nor rebuttal data)
Quality of content	Use of information from informative source texts Strong persuasiveness (but could be higher if there were counterarguments which were refuted by rebuttals) Good factual accuracy
General textual information	No introduction Conclusion No irrelevant information Average text length (250–400 words) Good writing mechanics

TABLE 11 | Elements of weak argumentative texts.

	Elements of weak argumentative texts	Number of texts (*n*)
Quality of argumentation	Only claim and arguments (no counterarguments nor rebuttals)	7
	Tunnel vision	7
Quality of content	Weak factual accuracy	4
	No use of information from source texts	5
	Weak persuasiveness	5
General textual information	Irrelevant information (= code 9)	5
	Short text length (<250 words)	4
	Bad writing mechanics	4

TABLE 12 | Elements of strong argumentative texts.

	Elements of strong argumentative texts	Number of texts (*n*)
Quality of argumentation	Use of (a) claim, (b) arguments, (c) counterclaim, (d) counterarguments, (e) rebuttal and (f) rebuttal data	4
	All counterarguments are refuted by rebuttal(s)	4
Quality of content	Use of information from source texts	7
	Strong persuasiveness	7
	Factual accuracy	6
General textual information	Introduction	4
	Conclusion	6
	Long text length (> 400 words)	7
	Good writing mechanics.	4

the rating procedures in the current study. More specifically, both the holistic and the analytical ratings were carried out by the same raters (see further in the section on limitations). The high attenuated correlation between absolute and comparative holistic rating was expected, as both are holistic procedures that look at the whole of the text. Alongside the high correlations between absolute holistic and absolute analytic rating and between absolute holistic and comparative holistic rating, the results in this study revealed rather low (but still significant at the 0.001 level) attenuated correlations between comparative holistic rating and absolute analytic rating. These results were expected given the different focus of the procedures (i.e., comparative holistic rating focuses on the whole text while analytic rating assesses different text features) and given the different underlying assessment strategy (i.e., analytic rating assesses texts in an absolute manner, while comparative holistic rating is based on comparing texts).

Based on the results of this study, the moderately to highly correlating rating procedures indicate the complexity of assessing argumentative texts. More specifically, argumentative writing is a complex interplay of various interrelated skills (such as reading skills, writing skills, and argumentation skills). Assessing such a complex and cognitively demanding activity requires assessment procedures that are able to grasp this complexity. The rating procedures central to this study each focused on assessing the quality of an argumentative text, resulting in relatively strong correlations, however they were far from perfectly aligned and, as the alluvial plot showed, the texts were ranked in different orders, which will be discussed in the next section.

RQ1b: How Often Do We See Deviations Between Rating Procedures and How Strong Are These Deviations?

Our findings showed that the rating procedures resulted in different ranking orders and that a text that is assigned a high score by one rating procedure, does not necessarily receive a high score by the other rating procedures. Given that the correlations are not fully aligned and as each rating procedure had its own focus of assessment (see RQ1a), this was expected. The deviations between the rating procedures were visualized in **Figure 1**. These findings reveal a certain level of agreement between the different procedures and indicate that despite different assigned scores, all procedures are suitable to assess the quality of an argumentative text.

We can conclude that the three rating procedures can be seen as proxies for the quality of argumentative texts, however, they have their own focus. Due to the nature of the analytic scoring process, the rating criteria in analytic rating are the most detailed. When all criteria are met, a high score is achieved, and although this is likely to result in high absolute and comparative holistic scores, this is not necessarily so. The opposite is true too: the best texts out of the comparative holistic approach might not necessarily have all elements required by the ASRAW. As all three procedures have their own focus, the scores will certainly not always be in line.

The conclusion that the texts are not exactly ranked in the same order by the three rating procedures should not necessarily be seen as a problem. It might be interesting to combine the different scores on one text, assigned by the different rating procedures, as feedback and input for the author. For example, as an author you can write a text that is assigned a low score by the ASRAW. An analytic rubric already offers opportunities for feedback: the author can clearly identify where points were lost (Bacha, 2001). But this same text could get a high score from comparative holistic rating (pairwise comparisons). The text then scores well in comparison to other texts written by peers. It can be interesting to look at texts written by peers: what can you learn from these texts in terms of writing mechanics, transitions between paragraphs, text length, text structure, etc.? In light of feedback, it therefore seems interesting to combine the input of different assessment procedures.

RQ2: Which Elements Characterize Mid-Range, Weak and Strong Texts?

The results indicate that certain text features or elements seem to be associated with mid-range (see **Table 10**), weak (see **Table 11**) or strong argumentative texts (see **Table 12**). In this discussion, we will elaborate on the text elements that can be decisive in judging a text as a strong argumentative text. Several studies have investigated the quality of argumentation in students' writing. In this respect, previous studies have pointed out that many students do not include counterarguments and rebuttals in their argumentative texts (Wolfe and Britt, 2008; Ferretti and Lewis, 2013). Very often students only include a claim and claim data from their own point of view, resulting in a tunnel vision in which the opposite view is ignored (Nussbaum and Kardash, 2005). Ideally, all viewpoints should be recognized and supported but the opposite viewpoint should be less convincing than the chosen viewpoint, as Stapleton and Wu (2015) declare. In the present study, we confirmed the results of several previous studies (Figueredo and Varnhagen, 2005; Barkaoui, 2010; De La Paz and Felton, 2010; Rezaei and Lovorn, 2010; Stapleton and Wu, 2015; Cuevas et al., 2016; Jansen et al., 2021; Syed et al., 2021). These studies showed that the elements that seem to be associated with strong texts were: (a) Use of the (adapted) Toulmin elements, (b) refuting all counterarguments by rebuttal(s), (c) integrating information from source texts, (d) strong persuasiveness, (e) factual accuracy, (f) use of introduction and conclusion, (g) long text length, and (h) good writing mechanics. If the integration of the abovementioned elements is related to the overall text quality, we need to teach students how to integrate these text elements in their argumentative writing, as Wong et al. (2008) suggest. Furthermore, it is also important to be aware of these essential genre elements when assessing argumentative texts (regardless of which rating procedure is used). In this respect, we need to inform raters of these success criteria. In absolute analytic rating this can be done by using a rubric in which these elements are present; in absolute and comparative holistic rating we can inform the raters of the key elements of a good argumentative text by means of training.

For Which Purposes Do Certain Procedures Work Well?

All three rating procedures each have their own advantages, a different focus and different prerequisites. In this section, we aim to guide practitioners, researchers and teachers in choosing a suitable rating procedure for the writing assignment they have in mind. Given the variation in scores, it is important to consider when to use which rating procedure. In what follows, we will discuss the purposes for which certain procedures work well. We briefly sum up the situations in which each rating procedure can be used and we provide advantages and disadvantages.

Absolute Holistic Rating Procedure

When in need of a quick general score, absolute holistic rating is ideal as this is a very time-efficient procedure (Charney, 1984). Scores can be assigned by one rater, making this procedure particularly useful for teachers and practitioners. However, raters ideally have some experience in rating texts (Charney, 1984; Rezaei and Lovorn, 2010). Holistic rating was used in our research to assess argumentative texts, but this procedure can be used for other text genres as well. A disadvantage of absolute holistic rating is that validity and reliability cannot be ensured (Wesdorp, 1981; Charney, 1984). The present study confirmed these previous studies as the absolute holistic rating procedure had a rather low reliability. However, this might be a problem for empirical researchers, but teachers and practitioners may value the quickness and naturalness of this procedure. In addition, we could address the low reliability by giving raters more guidance and training (Charney, 1984), e.g., in using the whole scoring range. In this respect, other absolute holistic assessments can also be implemented, e.g., a holistic rubric instead of general impression marking may help to obtain more reliable scores (Penny et al., 2000).

Comparative Holistic Rating Procedure (Pairwise Comparisons)

Pairwise comparisons use multiple raters to develop a rank order from lowest to highest text quality. Consequently, the need for multiple raters makes it difficult to implement this rating procedure in daily practice. However, as Bouwer et al. (2018) claim, assessing competences through pairwise comparisons is an easier task than using an analytic rubric which precisely pays attention to multiple text features. As high validity and reliability can be achieved, this procedure is very interesting for empirical researchers. Neither absolute holistic nor analytic rating automatically guarantee reliability, as we discussed above. A reliable rating procedure will, if applied again, obtain similar results in a following measurement (Charney, 1984). In our findings, we achieved an SSR of 0.83 for our pairwise comparisons. Researchers or practitioners that choose to use this procedure should pay attention to the provided instruction. It is possible that raters pay equal attention to quality of content, quality of argumentation, and general textual information. However, this cannot be fully assured: You cannot be sure in advance whether assessors will pay equal attention to these elements. Raters can always be influenced by their own thoughts on what defines a good text. Special attention should also be paid to text length, as our research demonstrated that longer texts were often rated as more qualitative texts. This may be due to the fact that text length is an easy, holistic criterion to use as this is the first visual indicator raters see when they are presented with two texts to compare (Lee et al., 2009).

Analytic Rating Procedure

In contrast to comparative holistic rating, analytic rating is workable for one person, making this a procedure that can be useful for teachers and practitioners. In addition, the analytic rating procedure can achieve high reliability (in our research: ICC = 0.98), but this is not automatically the case. Earlier research on reliability of analytic rating is still inconsistent. Therefore, Harsch and Martin (2013) suggest combining holistic and analytic rating procedures to achieve more reliable and valid results. In contradiction to the holistic rating procedure, training a rater is less important as raters only have to decide the category in which a certain text feature can be put, without further justification. However, we do not claim that analytic rating is always easy; deciding the level in which a certain feature belongs can be a difficult choice to make when there is doubt. In addition, raters need a clear view on the argumentative elements when analytically rating argumentative texts. Identifying claims or counterclaims is not self-explanatory. The absolute analytic rating procedure, and more specifically the ASRAW, can only be used when the main focus is on the structure of the argumentation. In this research, the ASRAW-rubric was used to assign scores to argumentative texts. Of course, other instruments can also be used to analytically score argumentative texts. The ASRAW mainly focuses on the quality of argumentation. If the structure of the argumentation is not good, the final score is automatically low. However, the ASRAW does pay attention to the quality of content, but only after taking a closer look at the structure of the argumentation. Writing mechanics and text length are not included in the ASRAW-rubric and therefore seemed to have less impact here compared to pairwise comparisons.

Limitations and Suggestions for Further Research

In what follows, limitations regarding the implementation of the rating procedures are addressed. In addition, suggestions for further research are proposed.

A first limitation focuses on the low reliability (ICC = 0.48) of the holistic rating procedure. We used general impression marking as the absolute holistic rating procedure, but it could have been interesting to use other absolute holistic assessment methods (e.g., holistic rubrics) as they provide additional resources to raters to assign a score to a text. As Weigle (2002) points out: "the scoring procedures are critical because the score is ultimately what will be used in making decisions and inferences about writers" (p. 108). Therefore, other assessment methods within holistic rating are also a possibility. We cannot guarantee that using a different holistic rating procedure would have had a positive effect on the ICC, but research by Penny et al. (2000) indicates that higher rater agreement could be achieved by means of using a holistic procedure containing additional support for raters.

The second limitation relates to the implementation of the rating procedures. Two out of three rating procedures (i.e., absolute holistic and analytic rating) were conducted by the first author and a second trained rater. Both raters rated the same texts holistically as well as analytically, which could partly explain correlations between the two procedures. The first author and the trained rater first implemented the holistic rating procedure, followed by the analytic rating procedure. This could have influenced the assessments, however, there was 1 week in between the ratings and the analytic ratings were implemented in a different order than the holistic ratings to avoid interdependency. For future research studies, we recommend that raters do not rate the same text holistically as well as analytically. Regarding validity and reliability, Harsch and Martin (2013) prefer rating a text both holistically and analytically. In providing feedback to students, this could be very useful. However, research by Hunter (1996) and Lee et al. (2009) showed that holistic and analytic scores are not always strongly correlated. More research is needed into the implications of merging information from both holistic rating and analytical rating. In our opinion, using both the holistic and the analytic rating procedure can indeed be a suitable way to assess texts, as Harsch and Martin (2013) suggest, but in practice it may not always be time-efficient and manageable to apply multiple rating procedures.

A third limitation focuses on the training of the raters. For holistic and analytic rating, both raters were intensively trained, unlike the 133 university students that conducted pairwise comparisons. The university students received a very short briefing, but no extensive training like the two raters that rated analytically and holistically was provided. The university students were no experts in assessing argumentative writing skills. Given the comparative nature of the writing assessment in the pairwise comparisons, we opted not to interfere so the university students could rely on their overall knowledge of argumentative writing, based on the provided broad criteria. Therefore we gave only few instructions to the university students. We notice some discrepancies in the educational literature on rating procedures. On the one hand, more general assessment studies of Sadler (1989) and Pollitt (2012) suggest that experienced raters can assess texts more easily, because of their experience. On the other hand, recent research on pairwise comparisons suggests that comparing texts is a relatively simple task and that rater experience is therefore not necessary (van Daal et al., 2016; Coertjens et al., 2017). On this view, pairwise comparisons may also work without an extended training. From this, we can conclude that training raters could have an influence on differences between the rating procedures, but for pairwise comparisons, little training should be sufficient in order to get reliable results. In addition, differences between raters cannot always be solved by training them in advance (Coertjens et al., 2017).

A fourth limitation relates to the 27 texts that were omitted from the study as they were informative instead of argumentative. These texts did not take a position (e.g., pro or contra), did not have the goal to persuade, nor were any arguments integrated. While this was a deliberate decision for research purposes, it is, however, not feasible in practice, as teachers cannot omit texts from evaluation. In a classroom context, all texts (whether argumentative or not) should be evaluated by the teacher.

CONCLUSION

The research field on writing assessment generally distinguishes between holistic and analytic rating procedures. However, another distinction has been recently identified: Absolute and comparative rating procedures (Bouwer and Koster, 2016; Coertjens et al., 2017). To date, there is little research that focuses on both distinctions. Therefore, this study is one of the first studies comparing absolute holistic rating, with comparative holistic rating (pairwise comparisons) and absolute analytic rating. In this study, we especially focus on the more innovative approach of pairwise comparisons, as this procedure is compared to more established methods of absolute holistic and analytic rating. In this study, it was indicated that the three rating procedures correlate moderately to highly, but each have different qualities, advantages and prerequisites. However, all three procedures are suitable for practitioners to use when assessing argumentative texts. In addition, we focused in detail on the deviance between the three rating procedures and the characteristics of mid-range, weak, and strong argumentative texts.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Faculty of Psychology and Educational Sciences of Ghent University (Specific Ethical Protocol for Scientific Research). Written informed consent for participation was not provided by the participants' legal guardians/next of kin because: There was an active written informed consent from the participants and a passive written informed consent from participants' parents.

AUTHOR CONTRIBUTIONS

YL, FD, BD, and HV designed the study. YL was in charge of the data collection procedure. YL analyzed the data, with the help of FD, BD, and HV. All authors wrote and reviewed the manuscript and approved its final version.

FUNDING

This research was supported by a grant from the Research Foundation Flanders (FWO) (Grant No. G010719N).

REFERENCES

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System* 29, 371–383. doi: 10.1016/S0346-251X(01)00025-2

Barkaoui, K. (2010). Explaining ESL essay holistic scores: a multilevel modeling approach. *Lang. Test.* 27, 515–535. doi: 10.1177/0265532210368717

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assess. Edu. Princ. Policy Pract.* 18, 279–293. doi: 10.1080/0969594X.2010.526585

Bouwer, R., and Koster, M. (2016). *Bringing research into the classroom: The effectiveness of Tekster, a newly developed writing program for elementary students*. Utrecht: Universiteit van Utrecht.

Bouwer, R., Goossens, M., Mortier, A. V., Lesterhuis, M., and De Maeyer, S. (2018). *Een comparatieve aanpak voor peer assessment: leren door te vergelijken. Toetsrevolutie: Naar Een Feedbackcultuur in Het Hoger Onderwijs*. Culemborg: Uitgeverij Phronese. 92–106.

Bramley, T., Bell, J. F., and Pollitt, A. (1998). Assessing changes in standards over time using Thurstone Paired Comparisons. *Edu. Res. Persp.* 25, 1–24.

Carr, N. T. (2000). A comparison of the effects of analytic and holistic rating scale types in the context of composition tests. *Issu. Appl. Ling.* 11:35. doi: 10.5070/l4112005035

Clark, D., and Sampson, V. (2007). Personally-seeded discussions to scaffold online argumentation. *Int. J. Educ. Sci.* 3, 351–361. doi: 10.1080/09500690600560944

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: a critical overview. *Res. Teach. Eng.* 18, 65–81.

Coertjens, L., Lesterhuis, M., Verhavert, S., Van Gasse, R., and De Maeyer, S. (2017). Teksten beoordelen met criterialijsten of *via* paarsgewijze vergelijking: Een afweging van betrouwbaarheid en tijdsinvestering. *Pedagogische Studien* 94, 283–303.

Cuevas, I., Mateos, M., Martín, E., Luna, M., and Martín, A. (2016). Collaborative writing of an argumentative synthesis from multiple sources: the role of writing beliefs and strategies to deal with controversy. *J. Writ. Res.* 8, 205–226. doi: 10.17239/jowr-2016.08.02.02

De La Paz, S., and Felton, M. K. (2010). Reading and writing from multiple source documents in history: effects of strategy instruction with low to average high school writers. *Contemp. Edu. Psychol.* 35, 174–192. doi: 10.1016/j.cedpsych.2010.03.001

Ferretti, R. P., and Lewis, W. E. (2013). "Best practices in teaching argumentative writing," in *Best practices in writing instruction* 2nd ed, eds S. Graham, C. A. MacArthur, and J. Fitzgerald (New York, NY: Guilford Press), 113–140.

Figueredo, L., and Varnhagen, C. K. (2005). Didn't you run the spell checker? Effects of type of spelling error and use of a spell checker on perceptions of the author. *Read. Psychol.* 26, 441–458. doi: 10.1080/02702710500400495

Follman, J. C., Anderson, J. A., and Anderson, J. A. (1967). An investigation of the reliability of five procedures for grading English themes. *Res. Teach. Eng.* 1, 190–200. doi: 10.1111/j.1365-2214.2011.01355.x

Freedman, S. (1981). Influences on evaluators of expository essays: beyond the text. *Res. Teach. Eng.* 15, 245–255.

Gill, T., and Bramley, T. (2013). How accurate are examiners' holistic judgements of script quality? *Assess. Edu. Princip. Policy Pract.* 20, 308–324. doi: 10.1080/0969594X.2013.779229

Goulden, N. R. (1994). Relationship of analytic and holistic methods to raters' scores for speeches. *J. Res. Dev. Edu.* 27, 73–82.

Graham, S., and Perin, D. (2007). What we know, what we still need to know: teaching adolescents to write. *Sci. Stud. Read.* 11, 313–335. doi: 10.1080/10888430701530664

Granado-Peinado, M., Mateos, M., Martín, E., and Cuevas, I. (2019). Teaching to write collaborative argumentative syntheses in higher education. *Read. Writ.* 32, 2037–2058. doi: 10.1007/s11145-019-09939-6

Harsch, C., and Martin, G. (2013). Comparing holistic and analytic scoring methods: issues of validity and reliability. *Assess. Edu. Princip. Policy Pract.* 20, 281–307. doi: 10.1080/0969594X.2012.742422

Hunter, D. (1996). The use of holistic versus analytic scoring for large-scale assessment of writing. *Canad. J. Prog. Eval.* 11, 61–85.

Huot, B. (1990). Reliability, validity, and holistic scoring: what we know and what we need to know. *Colleg. Comp. Commun.* 41:201. doi: 10.2307/358160

Huot, B. A. (1993). "The influence of holistic scoring procedures on reading and rating student essays," in *Validating holistic scoring for writing assessment: theoretical and empirical foundations*, eds M. M. Williamson and B. A. Huot (New York, NY: Hampton Press).

Jansen, T., Vögelin, C., Machts, N., Keller, S., and Möller, J. (2021). Don't just judge the spelling! the influence of spelling on assessing second-language student essays. *Front. Learn. Res.* 9:44–65. doi: 10.14786/flr.v9i1.541

Jarvis, S., Grant, L., Bikowski, D., and Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *J. Second Lang. Writ.* 12, 377–403. doi: 10.1016/j.jslw.2003.09.001

Kibler, A. K., and Hardigree, C. (2017). Using Evidence in l2 argumentative writing: a longitudinal case study across high school and university. *Lang. Learn.* 67, 75–109. doi: 10.1111/lang.12198

Krippendorff, K., and Hayes, A. F. (2007). Answering the call for a standard reliability measure for coding data. *Commun. Methods Measur.* 1, 77–89. doi: 10.1080/19312450709336664

Lee, Y. W., Gentile, C., and Kantor, R. (2009). Toward automated multi-trait scoring of essays: investigating links among holistic, analytic, and text feature scores. *Appl. Ling.* 31, 391–417. doi: 10.1093/applin/amp040

Lord, F. M., and Novick, M. R. (1968). *Statistical theories of mental test scores*. Boston, MA: Addison-Welsley.

Marttunen, M., Laurinen, L., Litosseliti, L., and Lund, K. (2005). Argumentation skills as prerequisites for collaborative learning among finnish, french, and english secondary school students. *Edu. Res. Eval.* 11, 365–384. doi: 10.1080/13803610500110588

McMahon, S., and Jones, I. (2015). A comparative judgement approach to teacher assessment. *Assess. Edu. Princip. Policy Pract.* 22, 368–389. doi: 10.1080/0969594x.2014.978839

Messick, S. (1989). Meaning and values in test validation: the science and ethics of assessment. *Edu. Res.* 18, 5–11. doi: 10.3102/0013189X018002005

Myers, M. (1980). *A Procedure for Writing Assessment and Holistic Scoring. In College Composition and Communication*. Urbana, IL: National Council of Teachers of English and Educational Resources Infor- mation Center

NCES (2012). *The Nation's Report Card: Writing* 2011. Washington, DC: NCES.

Nussbaum, E. M., and Kardash, C. M. (2005). The effects of goal instructions and text on the generation of counterarguments during writing. *J. Edu. Psychol.* 97, 157–169. doi: 10.1037/0022-0663.97.2.157

Nussbaum, E. M., and Schraw, G. (2007). Promoting argument-counterargument integration in students' writing. *J. Exp. Edu.* 76, 59–92. doi: 10.3200/JEXE.76.1.59-92

Penny, J., Johnson, R. L., and Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability: an empirical study of a holistic rubric. *Assess. Writ.* 7, 143–164. doi: 10.1016/S1075-2935(00)00012-X

Pollitt, A. (2012). Comparative judgement for assessment. *Int. J. Technol. Design Edu.* 22, 157–170. doi: 10.1007/s10798-011-9189-x

Qin, J., and Karabacak, E. (2010). The analysis of Toulmin elements in Chinese EFL university argumentative writing. *System* 38, 444–456. doi: 10.1016/j.system.2010.06.012

Rezaei, A. R., and Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assess. Writ.* 15, 18–39. doi: 10.1016/j.asw.2010.01.003

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instruct. Sci.* 18, 119–144. doi: 10.1007/BF00117714

Sasaki, M., and Hirose, K. (1999). Development of an analytic rating scale for Japanese L1 writing. *Lang. Test* 16, 457–478. doi: 10.1177/026553229901600403

Simon, S. (2008). Using Toulmin's Argument Pattern in the evaluation of argumentation in school science. *Int. J. Res. Method Edu.* 31, 277–289. doi: 10.1080/17437270802417176

Song, Y., and Ferretti, R. P. (2013). Teaching critical questions about argumentation through the revising process: effects of strategy instruction on college students' argumentative essays. *Read. Writ.* 26, 67–90. doi: 10.1007/s11145-012-9381-8

Stapleton, P., and Wu, Y. (2015). Assessing the quality of arguments in students' persuasive writing: a case study analyzing the relationship between surface structure and substance. *J. Eng. Acad. Purp.* 17, 12–23. doi: 10.1016/j.jeap.2014.11.006

Syed, S., Al-Khatib, K., Alshomary, M., Wachsmuth, H., and Potthast, M. (2021). Generating informative conclusions for argumentative texts. *arXiv* doi: 10.48550/arXiv.2106.01064

Thorndike, E. (1920). A constant error in psychological ratings. *J. Appl. Psychol.* 4, 25–29. doi: 10.1037/h0071663

Thurstone, L. L. (1927). A law of comparative judgment. *Psychol. Rev.* 34, 273–286. doi: 10.1037/0033-295X.101.2.266

Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Traga Philippakos, Z. A., and MacArthur, C. A. (2019). Integrating collaborative reasoning and strategy instruction to improve second graders' opinion writing. *Read. Writ. Quart.* 2019, 1–17. doi: 10.1080/10573569.2019.1650315

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., and De Maeyer, S. (2016). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assess. Edu. Princip. Policy Pract.* 26, 59–74. doi: 10.1080/0969594X.2016.1253542

van Rijt, J. H. M., van den Broek, B., and De Maeyer, S. (2021). Syntactic predictors for text quality in Dutch upper-secondary school students' L1 argumentative writing. *Read. Writ.* 34, 449–465. doi: 10.1007/s11145-020-10079-5

Varghese, S. A., and Abraham, S. A. (1998). Undergraduates arguing a case. *J. Second Lang. Writ.* 7, 287–306. doi: 10.1016/S1060-3743(98)90018-2

Veal, R. L., and Hudson, S. A. (1983). Direct and indirect measures for large-scale evaluation of writing. *Res. Teach. Eng.* 17, 290–296. doi: 10.3390/v13081651

Verhavert, S., De Maeyer, S., Donche, V., and Coertjens, L. (2018). Scale separation reliability: what does it mean in the context of comparative judgment? *Appl. Psychol. Measur.* 42, 428–445. doi: 10.1177/0146621617748321

Vögelin, C., Jansen, T., Keller, S. D., Machts, N., and Möller, J. (2019). The influence of lexical features on teacher judgements of ESL argumentative essays. *Assess. Writ.* 39, 50–63. doi: 10.1016/j.asw.2018.12.003

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Wesdorp, H. (1981). *Evaluatietechnieken voor het moedertaalonderwijs*. Den Haag: Staatsuitgeverij.

Wolfe, C. R., and Britt, M. A. (2008). The locus of the myside bias in written argumentation. *Think. Reason.* 14, 1–27. doi: 10.1080/13546780701527674

Wolfe, E. W., Song, T., and Jiao, H. (2016). Features of difficult-to-score essays. *Assess. Writ.* 27, 1–10. doi: 10.1016/j.asw.2015.06.002

Wong, B. Y. L., Hoskyn, M., Jai, D., Ellis, P., and Watson, K. (2008). The comparative efficacy of two approaches to teaching sixth graders opinion essay writing. *Contemp. Edu. Psychol.* 33, 757–784. doi: 10.1016/j.cedpsych.2007.12.004

Yune, S. J., Lee, S. Y., Im, S. J., Kam, B. S., and Baek, S. Y. (2018). Holistic rubric vs. analytic rubric for measuring clinical performance levels in medical students. *BMC Med. Edu.* 18:1–6. doi: 10.1186/s12909-018-1228-9

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 13 May 2022
doi: 10.3389/feduc.2022.823895

Validity of Comparative Judgment Scores: How Assessors Evaluate Aspects of Text Quality When Comparing Argumentative Texts

Marije Lesterhuis[1]*, *Renske Bouwer*[2], *Tine van Daal*[1], *Vincent Donche*[1] and *Sven De Maeyer*[1]

[1] *Edubron, Department of Training and Educational Sciences, University of Antwerp, Antwerp, Belgium,* [2] *Utrecht Institute of Linguistics OTS, Utrecht University, Utrecht, Netherlands*

OPEN ACCESS

Edited by:
Anders Jönsson,
Kristianstad University, Sweden

Reviewed by:
Stephen Humphry,
University of Western Australia, Australia
Wei Shin Leong,
Ministry of Education, Singapore

***Correspondence:**
Marije Lesterhuis
Marije.Lesterhuis@uantwerpen.be

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *28 November 2021*
Accepted: *21 March 2022*
Published: *13 May 2022*

Citation:
Lesterhuis M, Bouwer R, van Daal T, Donche V and De Maeyer S (2022) Validity of Comparative Judgment Scores: How Assessors Evaluate Aspects of Text Quality When Comparing Argumentative Texts. Front. Educ. 7:823895. doi: 10.3389/feduc.2022.823895

The advantage of comparative judgment is that it is particularly suited to assess multidimensional and complex constructs as text quality. This is because assessors are asked to compare texts holistically and to make a quality judgment for each text in a pairwise comparison based upon on the most salient and critical differences. Also, the resulted rank order is based on the judgment of all assessors, representing the shared consensus. In order to be able to select the right number of assessors, the question is to what extent the conceptualization of assessors prevails in the aspects they base their judgment on, or whether comparative judgment minimizes the differences between assessors. In other words, can we detect types of assessors who tend to consider certain aspects of text quality more often than others? A total of 64 assessors compared argumentative texts, after which they provided decision statements on what aspects of text quality had informed their judgment. These decision statements were coded on six overarching themes of text quality: argumentation, organization, language use, language conventions, source use, references, and layout. Using a multilevel-latent class analysis, four different types of assessors could be distinguished: narrowly focused, broadly focused, source-focused, and language-focused. However, the analysis also showed that all assessor types mainly focused on argumentation and organization, and that assessor types only partly explained whether the aspect of text quality was mentioned in a decision statement. We conclude that comparative judgment is a strong method for comparing complex constructs like text quality. First, because the rank order combines different views on text quality, but foremost because the method of comparative judgment minimizes differences between assessors.

Keywords: comparative judgment, validity, writing assessment, assessor cognition, latent class analysis

INTRODUCTION

Comparative judgment is particularly suited for the judgment of complex skills, competencies and performances, such as writing or mathematical problem solving (Pollitt and Crisp, 2004; Heldsinger and Humphry, 2010; Pollitt, 2012; Jones et al., 2015). A characteristic of the assessment of complex skills is that the quality of students' work cannot be considered as either right or wrong, but on a continuum of quality. The quality is determined by multiple aspects that are highly intertwined (Sadler, 2009). For text quality this is for example the content, structure, style and grammar. To judge the quality, human judgment is key. However, it is important that scores reflect the complexity of the skill under assessment. Therefore, the way assessors come to a judgment within comparative judgment plays a major role in its validity argument (Bejar, 2012).

Within comparative judgment, assessors base their judgment on holistic interpretations of the quality of students' work. That means that the assessor makes a single comparison of which text is better considering the communicative effectiveness. In other words, to what extent is a text reaching its communicative goal? These holistic interpretations take into account the complexity of what quality comprises (Lesterhuis et al., 2019). Also, when comparing, assessors can rely on their expertise and their own conceptualization of quality. This supports the validity of the assessment scores, because final scores are based on all judgments made by all assessors and thus represent the shared consensus on what quality comprises (Jones and Inglis, 2015). Previous empirical studies have indeed shown that assessors focus on different aspects when they make comparisons. Hence, involving a group of assessors in comparative judgment enhances construct representation (Pollitt and Whitehouse, 2012; Whitehouse, 2012; van Daal et al., 2019).

Yet, there is still little insight in the number and type of assessors that should be involved for valid comparative judgments and the role assessors play to achieve full construct representation (Messick, 1989; Lesterhuis et al., 2019). Therefore, we need to know the extent to which assessors differ from each other regarding the probability that certain aspects of quality are assessed. Up to now, studies only focused upon differences between assessors (van Daal et al., 2019) but didn't look for different profiles or types of assessors. Therefore, this study investigates whether assessors focus on different or similar aspects of students' argumentative texts when making comparative judgments, and to what extent different types of assessors can be distinguished based on their judgments. These insights help future assessment coordinators with the selection of assessors in order to achieve text scores that can be validity interpreted as representing the quality of the texts.

BACKGROUND

To understand the role that assessors play in a valid interpretation of text scores, this section discusses how assessors make comparative judgments, the types of assessors that can be distinguished with respect to the aspects of text quality they value and previous studies within the field of comparative judgment that looked into the aspects assessors base their decisions on.

Assessors' Judgment Process

The assessment of text quality requires the assessors to translate the text' quality into a judgment on that quality. In case of comparative judgment this is the decision which text is of higher quality. Therefore, an assessor reads the two texts, interprets the texts considering the different aspects and formulates a judgment on the quality of the whole. Assessors' cognition or mental scheme determines the way the assessors go through the text and how they conceptualize the texts' quality. Especially the latter is important for a valid interpretation of text scores because it affects what kind of aspects assessor do and do not value when assessing text quality. Consequently, the results of pairwise comparisons are based on assessors' conceptualization of text quality.

However, assessors can differ in how they conceptualize text quality (Huot, 1990; Vaughan, 1991). Various studies that look into how assessors judge single texts have investigated possible causes of this difference. For example, Cummings and others found that second language raters pay more attention to language than to rhetoric ideas, in contrast to first language raters. Wolfe (1997) found that proficient raters focus more on general features. Wang et al. (2017) found that inexperienced assessors differ in how they consider textual borrowing, the development of ideas, and the consistency of the focus. Consequently, all these studies show that assessors differ in how they translate texts into scores. Therefore, key in the validity argument is understanding how within a scoring method as comparative judgments, the selection of assessors affects construct representation.

Differences Between Assessors and Assessor Types

Studies on the assessment of text quality based on holistic and analytic scoring show that the aspects assessors consider is not fully random, but that assessors tend to belong to a certain type. For instance, Diederich et al. (1961) asked assessors to score 300 texts holistically on a nine-point scale without any instructions or criteria; meanwhile, the assessors had to provide the texts with written comments. The assessors differed to a large extent regarding the quality level to which they assigned the texts. Based on these differences, the researchers classified the assessors into five groups. Additionally, the researchers analyzed the assessors' comments and examined whether the comments differed between the groups. All assessors had focused on the clarity of expression, coherence, and logic (reasoning). However, the groups differed in the importance they attached to the relevance, clarity, quantity, development, and soundness of ideas (idea-focused); on organization and spelling (form-focused); on style, interest and sincerity (creativity-focused), on the errors in texts (mechanics-focused); on the choice and arrangement of words (effectiveness-focused). Based on this study, we can expect that assessors differ in what aspects of text they value while comparing two texts and consequently also the way they score text quality.

To look into the effect of conceptualization of text quality on analytic scoring with criteria, Eckes (2008) analyzed the importance that 64 experienced assessors attributed to nine quality criteria. He identified six groups of assessors. Four groups were more-or-less like the groups identified by Diederich et al. (1961); the groups focused on syntax, correctness, structure, and fluency. However, Eckes also found two groups that could be typified according to the aspects they considered less important compared to the other assessor types: not fluency-focused and not argumentation-focused. In a follow-up study, Eckes (2012) showed that the groups were related to how these assessors scored texts on a rating scale. He found that belonging to a certain type of assessors relates to how severe an assessor rates a criterion.

Using a similar approach, Schaefer (2016) found three groups of assessors when analyzing how 40 relatively untrained English teachers scored 40 English essays using criteria. He distinguished the assessors that focused on rhetorical features, linguistic features and mechanics. Schaefer (2016) could, however, not substantiate the link between the aspects the assessors said they valued and the aspects they really valued when scoring texts. Nevertheless, these studies showed that assessors differ in how they conceptualize text quality and that this affects how outcomes, in this case scores, using holistic or criteria-based scoring methods.

Differences Between Assessors in Comparative Judgment

A main advantage of comparative judgment is that assessors only have to provide relative decisions. Consequently, differences in severity (i.e., one assessor systematically giving lower scores) do not affect the reliability of the results anymore. Most studies indeed show that the resulting rank order of the comparative judgments of multiple assessors is highly reliable. Using 10 till 14 comparisons per text (or other types of student work) generates already acceptable reliability estimates (Separation Scale Reliability = SSR) of 0.70 (Verhavert et al., 2019). A high SSR reflects a high stability in the way the texts are ranked (Verhavert et al., 2018). This is a prerequisite for valid scores. However, valid scores also require that these quality scores fully reflect the complex construct of text quality.

To what extent do assessors take the full construct of text quality in account when making comparative judgment? Some empirical studies have already investigated the aspects that assessors consider when choosing texts by looking into how assessors justify their decision. In the study of van Daal et al. (2019), the explanations of 11 assessors to justify their decisions when comparing academic papers were analyzed. Analysis of these explanations—or decision statements—revealed that the group of assessors considered all relevant aspects of text quality and did not consider irrelevant aspects as the basis of their comparative judgments. Also, all assessors focused predominantly on the structure of the text and source use. However, there was still considerable variance between assessors. They varied in the aspects they also considered and in the number of aspects they mentioned in their decision statements. For example, some assessors focused on the discussion section, while others did not, and some mentioned language errors, while others did not. In the study of Lesterhuis et al. (2019), 27 teachers compared argumentative texts, referring to a wide range of aspects of text quality when justifying their decisions, varying from aspects of the argumentation to whether a title was present. However, whether assessors differ systematically on the aspects they discriminate on when comparing texts, or whether this has been caused by the different text pairs has not been established.

Humphry and Heldsinger (2019) show that the texts that are in a certain pair inform what assessors consider. They asked assessors to tick which aspects of 10 criteria informed their judgment after making a comparison. They found that when assessors compare lower quality texts, assessors more often base their decision on sentence structure and spelling and grammar and when comparing texts of higher quality, they referred to audience orientation and setting and character. This raises the question whether there are trends among assessors in the aspects they consider when comparing texts, independent of the pair of texts. In other words, can different types of assessors be detected when looking at the aspects assessors refer to when justifying their decision?

RESEARCH AIMS

Previous research focusing on other scoring methods has shown that assessors develop different conceptualizations of text quality which affect how they judge the quality of texts. It is yet unknown whether assessors take different aspects into account when making a comparison decision. This is relevant because in the method of comparative judgments, assessors can play a major role in the aspects that are considered because they are not forced to assess text quality on predefined quality criteria (as is the case in analytic judgments), but instead they can rely on their own expertise when comparing texts in a holistic manner (e.g., Pollitt, 2012; Jones et al., 2015). Previous studies already suggest that assessors make comparative judgments based on a wide range of relevant aspects of text quality, showing that the shared consensus of the resulting rank-order reflects the complexity of the construct of text quality (van Daal et al., 2017; Lesterhuis et al., 2019). This does not fully reveal whether assessors are comparable or whether different types of assessors exist, and hence, multiple assessors are needed for a valid assessment. Therefore, the central question in this study is whether different types of assessors can be distinguished that tend to base their comparative judgments on certain aspects. And when types of assessors can be distinguished, how can we typify these classes? These insights are important to understand the role of assessors in the validity argument and how the selection of assessors affects a valid interpretation of text scores.

MATERIALS AND METHODS

Participants

To search for trends among assessors, we chose a varied selection of assessors. A total of 64 assessors participated in this study. They had an average age of 37.23 years (*SD* = 14.22), 20 were men, 44 were women, and all were native Dutch speakers. Of the 64 assessors, 32.8% were student teachers, with no experience teaching or evaluating students' work; 42.2% were teachers (years of experience *M* = 19.96, *SD* = 13); 14.1% were teacher trainers (years of experience *M* = 13.11, *SD* = 7.67); and 9.4% worked as examiners (years of experience *M* = 23, *SD* = 9.17) working for an organization that certifies students who are following an irregular educational track.

The Assessment

The assessors evaluated the quality of three argumentative writing tasks completed by 135 students at the end of secondary education in their first language (Dutch). The students had to write an argumentative essay about the following topics: "Having children," "Organ donation," and "Stress at school." These tasks were previously used in the research of van Weijen (2009) but were adjusted slightly to the Flemish context. For each task, the students received six short sources, which they had to use to support their arguments. We included three tasks, so the findings do not depend on one specific task.

The tasks were in line with the competence "argumentative writing" as formulated in the final attainment goals of the Flemish Department of Education.[1] These goals were familiar to all assessors and students, and described what students need to be able to at the end of secondary education. The students had 25 min for each task. The 135 texts with the topics "Having children" and "Organ donation" were used. However, due to practical issues, only 35 randomly selected texts with the topic "Stress at school" were included in this study.

Procedure

Assessors came together on the campus, two times for 2 h. Before starting the assessment, the assessors received an explanation about the method of comparative judgment. Also, we gave a short introduction of the students' tasks and the competence of argumentative writing. The Digital Platform for the Assessment of Competences tool (D-PAC) supported the assessments that used comparative judgment. Within this tool, three assessments were created, each including texts of only one topic. For the topic "Having children," 1,224 pairs were generated; for the topic "Organ donation," 901 pairs were generated; and for the topic "Stress at school," 474 pairs were generated. In total, 2,599 comparative judgments were made. These pairs were randomly assigned to the assessors, who started with the assessment of "Having children," followed by "Organ donation" and "Stress at school." For each pair, the assessors decided which of the two texts was of higher quality in light of the competence of argumentative writing.

Next, they responded after each comparison to the query "Can you briefly explain your judgment?" Based on these decision statements, information was gathered on the aspects of text quality that informed the decisions of the assessors (Whitehouse, 2012; Bartholomew et al., 2018; van Daal et al., 2019). Each assessor made at least 10 comparisons, with a maximum of 56 comparisons (M = 40.60, SD = 16.16). The variation in the number of comparisons was due to assessors only attended one judgment session and/or because of differences in judgment speed. The assessors provided a decision statement for 98,1% of the made comparisons.

Pre-analysis

Using user-defined functions in R, we applied the Bradley-Terry-Luce model to the data (Bradley and Terry, 1952; Luce, 1959), in order to estimate logit scores for each text. These logit scores express the log of the odds, which can be transformed to the probability that a particular text will be selected as the better text when compared to a text of average quality. These scores can be interpreted as a text quality score. The reliability of these scores was calculated by taking the variation in quality and the standard error of each text's quality score. This reliability is expressed in the scale separation reliability (*SSR*). The texts with the topic "Having children" had an *SSR* of 0.81, "Organ donation" was 0.73 and "Stress at school" 0.89. These high reliabilities show that the comparative judgments across the assessors were consistent (Verhavert et al., 2019).

Analyses

All decision statements were coded according to seven aspects of text quality, argumentation, organization, language use, formal language conventions, source use and references. A total of 10% of the assessment "having children" was double coded and showed a sufficient level of reliability of K = 0.65 (Stemler, 2004). **Table 1** shows the percentage each element was mentioned according to all the assessments and assessors.

In order to detect whether types of assessors can be distinguished, a data file was created in which it was indicated for each comparison whether the assessor had mentioned an aspect of text quality (1) or not (0). A multilevel latent class (MLCA) analysis was performed on this dataset, as comparisons were nested in assessors. A latent class analysis investigates if there are trends in the answers given by assessors, by examining the probability of an aspect being mentioned by an assessor. Assessors with the same probability of mentioning an aspect are grouped in a class (Vermunt and Magidson, 2003). By describing this class, a type of assessor is created.

In order to determine how many classes of assessor types can be distinguished, several models are estimated. Each model contains one class more than the previous model. To select the best fitting model, we looked first into the Bayes Information Criterion (*BIC*) and the total Bivariate Residuals (*TBVR*). For both the *BIC* and *TBVR*, we were interested in the relative reduction, which indicates the importance of adding another class to the class solution regarding model fit (van den Bergh et al., 2017).

Second, we used the classification error and entropy (E) to investigate the different class solutions. The classification error refers to the certainty that each assessor can be assigned to one of the distinguished classes. The classification error increases when several assessors show a high probability of belonging to more than one class. The entropy is a single number summary of the certainty with which assessors can be assigned to a class. This depends, on the one hand, on the overlap of classes with regard to their probability patterns and, on the other hand, on how well assessors can be assigned to a single class according to their modal posterior probabilities. The closer the entropy

[1] www.onderwijsdoelen.be

TABLE 1 | Coding scheme with example statements and percentages elements were mentioned.

Aspect of text quality	Example statement	% mentioned in decision statement (*N* = 2,599)
Argumentation	The arguments used in the left text are better supported (comparison 164, a teacher with 3 years of experience)	57.3
Organization	I think the organization and form of the structure are better (comparison 227, a teacher trainer with 6 years of experience)	56.0
Language use	Beautiful and surprising use of language is important when you aim to convince someone (comparison 359, a teacher with 7 years of experience)	23.4
Formal language conventions	However, this text has grammar and construction mistakes (comparison 1,977, a teacher with 30 years of experience)	19.2
References	The references to sources are done well (comparison 2,309, a student with no experience)	19.2
Source use	Better integration of the sources (comparison 1,713, a student with no experience)	18.2
Layout	. . . The aspects of the text quality that are seen easily, as layout, length and the presence of a title	17.0

is to 1, the more certain assessors can be assigned to class (Collins and Lanza, 2009).

When the best fitted model of classes is selected, the differences between classes will be described by a horizontal and a vertical analysis. This description results in different types of assessors. Using the Wald statistic, we examine whether a class differs significantly from other classes with regard to aspects that are mentioned (horizontal analysis). In addition, for each class we will look at the probability that particular text quality aspects are mentioned in the decision statements (vertical analysis).

Because experience and occupational background can be related to how assessors conceptualize text quality, we checked whether these assessors' characteristics were related to the class composition. Experience has been operationalized as the assessors' number of years of relevant experience of teaching and/or writing assessment. To check the relationship with the group composition, the Welch test was executed, because the number of years did not meet the assumption of equal variances between groups. The occupational background was operationalized as an assessor being a student teacher, teacher, teacher trainer or examiner. To check the relationship with group composition, a chi-square test was performed.

RESULTS

Exploring the Number of Assessor Classes

Table 2 shows that the *BIC* and the *TBVR* kept deteriorating by adding a class to the class solution. However, the relative increase of the model fit stopped after the four-class solution.

The four-class solution also appeared to be good when investigating the certainty that assessors could be assigned to a class. According to the classification error and the entropy presented in **Table 2**, the four-class solution resulted in a better assignment of assessors to classes than the two-, three-, or five-class solution, as the classification error was 0.02 and the entropy 0.96. To illustrate, when applying a four-class solution, 61 assessors can be assigned to a class with a probability exceeding 90%. For the remaining three assessors, the highest probability to belong to a class is 77, 71, and 59%. Based on these results, we argue that assessors can be divided into four homogeneous sub-classes concerning the probability that they refer to an aspect of text quality.

TABLE 2 | Model parameters and classification of assessors to classes.

Model	*BIC* (*LL*)	Δ *BIC* (*LL*)	*TBVR*	Classification error	*E*
One class	19,912.42	46.64	–	1	
Two classes	19,447.51	−464.91	36.34	0.02	0.93
Three classes	19,236.24	−211.27	30.91	0.02	0.95
Four classes	19,118.89	−117.36	27.62	0.02	0.96
Five classes	19,022.71	−96.17	24.74	0.03	0.95
Six classes	18,937.78	−84.93	22.26	0.01	0.97
Seven classes	18,863.06	−74.73	19.90	0.02	0.97
Eight classes	18,816.70	−46.36	18.08	0.03	0.96

Describing the Differences Between Assessor Classes

This section describes the class solution in greater depth. It begins with a general description of the four-class solution and is followed by a description of each of the four classes.

The Class Solution

The best class solution divided the 64 assessors into four assessor classes. The classes differed in size, however, each class consisted of a substantial number of assessors. The first class consisted of 35.06% (*n* = 22) of the assessors, the second class 32.52% (*n* = 21), the third class 18.74% (*n* = 12), and the fourth class 13.68% (*n* = 9).

The four classes differed significantly in each aspect of text quality that they mentioned, as shown by the Wald tests ($W \geq 52.95$, $p < 0.01$) and in the average number of aspects they mentioned in a decision statement [Welch's $F(3,1062.12) = 243.17, p < 0.01$]. The R^2 in **Table 3** shows that the extent that this class' solution explained whether an aspect of text quality was mentioned, varied between 0.02 for argumentation and 0.11 for the layout. In other words, although differences between the classes are significant, the class solution does not fully explain whether a particular aspect of text quality was mentioned in the decision statements.

TABLE 3 | Explanatory power of the four-class solution.

	R^2
Argumentation	0.02
Organization	0.08
Language use	0.06
Source use	0.09
Language conventions	0.03
References	0.07
Layout	0.11

Four Types

To describe the classes in depth, **Table 4** reflects the differences between the classes when using the Wald statistic with a paired comparison approach. **Figure 1** visualizes the probability an aspect of text quality was mentioned by each class.

Class 1 contained the largest number of assessors ($n = 22$) and can be indicated as language-focused. This class referred most often to the organization of texts, and subsequently to the argumentation. However, typical for this class is that it additionally referred regularly to language use and language conventions, with 36 and 26%, respectively. For language use, this probability is significantly higher than the other classes. For language conventions, the probability is higher than the assessors in class 2 and class 3. Moreover, this class referred to 2.43 ($SD = 1.13$) aspects of text quality in a decision statement, on average.

Class 2 ($n = 21$) can be called narrowly focused. Only argumentation and organization were deemed to be relevant to these assessors. However, with 48% for argumentation and 41% for organization, this class did not even refer to these aspects regularly, compared to the other classes. The narrow focus is also reflected in the number of aspects mentioned, on average, in each decision statement ($M = 1.38$, $SD = 0.94$). The Games-Howell *post hoc* test showed that this was significantly less than the other classes ($p < 0.01$).

Class 3 ($n = 12$) can be indicated as source-focused. Besides the argumentation and organization of texts, this class found source use and references to be the most important aspects for choosing a text. The 34% probability for source use is significantly higher than the other classes, and the 28% probability for references is significantly more than classes 1 and 2. This class reflected on 2.16 ($SD = 1.14$) aspects per decision statement on average.

Class 4 ($n = 9$) can be typified as broadly focused. In addition to argumentation, this class was more likely to refer to all aspects than at least two of the other classes. Moreover, each aspect was mentioned with more than a 25% probability. That broad focus was also reflected in 3.04 ($SD = 0.94$) aspects that this class averagely mentioned in a decision statement. According to the Games-Howell *post hoc* test, this is significantly more than

TABLE 4 | The probability that assessors *w*ithin a class refer to an aspect of text quality.

Class 1 language-focused		Class 2 narrowly focused	Class 3 source-focused	Class 4 broadly focused	Average probability
Argumentation	0.63	0.48*	0.66	0.55*	0.57
C2: $W = 36.42$, $p < 0.01$ C3: $W = 1.27$, $p = 0.2e6$ C4: $W = 6.43$, $p = 0.01$		C3: $W = 36.64$, $p < 0.01$ C4: $W = 4.07$, $p = 0.04$		C4: $W = 10.04$, $p < 0.01$	
Organization	0.71	0.41	0.46	0.71	0.56
C2: $W = 148.11$, $p < 0.01$ C3: $W = 79.74$, $p < 0.01$ C4: $W = 0.00$, $p = 0.96$		C3: $W = 2.61$, $p = 0.11$ C4: $W = 78.39$, $p < 0.01$		C4: $W = 48.10$, $p < 0.01$	
Language use	0.36*	0.13	0.15	0.28*	0.23
C2: $W = 113.64$, $p < 0.01$ C3: $W = 63.30$, $p < 0.01$ C4: $W = 7.25$, $p < 0.01$		C3: $W = 0.75$, $p = 0.39$ C4: $W = 33.53$, $p < 0.01$		C4: $W = 19.20$, $p < 0.01$	
Source use	0.16*	0.06*	0.34	0.33	0.18
C2: $W = 40.24$, $p < 0.01$ C3: $W = 54.03$, $p < 0.01$ C4: $W = 42.55$, $p < 0.01$		C3: $W = 144.21$, $p < 0.01$ C4: $W = 123.18$, $p < 0.01$		C4: $W = 0.03$, $p = 0.86$	
Language conventions	0.26	0.11*	0.15*	0.27	0.19
C2: $W = 56.60$, $p < 0.01$ C3: $W = 20.27$, $p < 0.01$ C4: $W = 0.08$, $p = 0.78$		C3: $W = 4.30$, $p = 0.03$ C4: $W = 40.15$, $p < 0.01$		C4: $W = 16.09$, $p < 0.01$	
References	0.16*	0.10*	0.28*	0.40*	0.19
C2: $W = 14.49$, $p < 0.01$ C3: $W = 25.83$, $p < 0.01$ C4: $W = 71.60$, $p < 0.01$		C3: $W = 67.41$, $p < 0.01$ C4: $W = 125.16$, $p < 0.01$		C4: $W = 12.75$, $p < 0.01$	
Layout	0.15	0.09*	0.12	0.49*	0.17
C2: $W = 11.82$, $p < 0.01$ C3: $W = 2.15$, $p = 0.14$ C4: $W = 131.85$, $p < 0.01$		C3: $W = 144.21$, $p < 0.01$ C4: $W = 123.18$, $p < 0.01$		C4: $W = 117.43$, $p < 0.01$	

**Significantly different from all other classes with $p < 0.05$.*

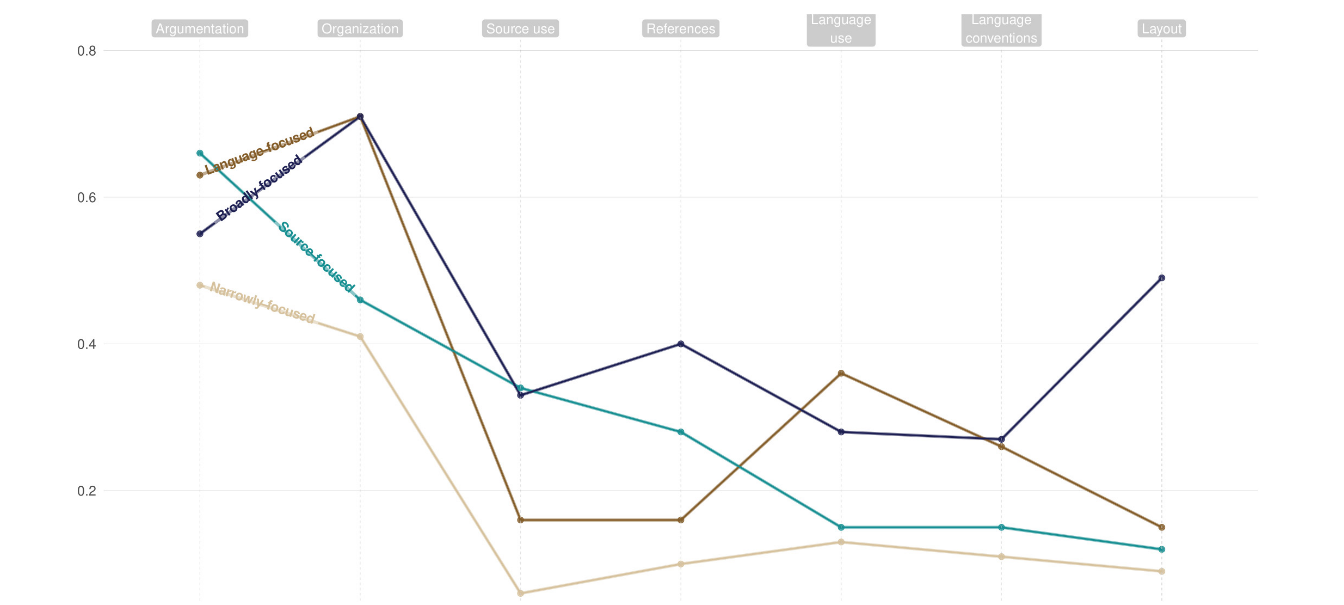


FIGURE 1 | The probability that a class referred to an aspect of text quality in a decision statement.

the average number of aspects mentioned by the other classes ($p < 0.01$).

Controlling for Experience

As the assessors differed in relevant years of experience and occupational background, we investigated whether these assessor characteristics related to the distinguished types. A Welch test showed that years of relevant experience had no significant effect on the composition of classes [$F(3, 25.84) = 0.91$, $p = 0.45$]. Next, the chi-square test showed there is no statistically significant relationship between occupational background and the classes [χ^2 (12, $N = 63$) = 9.31, $p = 0.68$].

DISCUSSION

Comparative judgment is especially suited to assess complex skills. As assessors are assumed to vary in the aspects upon which they focus, combining their judgments should foster construct representation (Pollitt and Whitehouse, 2012; Whitehouse, 2012; van Daal et al., 2019). However, it is unclear whether differences between assessors occur systematically. Therefore, this study examined to what extent types of assessors can be discerned. A type of assessor refers to a group of assessors that systematically considers an aspect of text quality (or not) when discerning between two texts. To investigate whether different types of assessors could be distinguished we analyzed 2,599 decision statements that 64 assessors gave to explain their comparative judgments on the quality of argumentative texts of students in the fifth grade of secondary education. These decision statements were coded on argumentation, organization, language use, source use, language conventions, references, and layout. Next, we applied a MLCA to investigate whether classes of assessors with a similar argumentation pattern could be detected.

Based on the MLCA, four classes of assessors could be distinguished. All assessor classes referred to organization and argumentation when making comparative judgments but differed with regards to other aspects of text quality. Class 1 was mainly language-focused. These assessors were more likely to mention language use and conventions to justify their comparative judgments than the other classes of assessors. Class 2 was narrowly focused, which means that assessors in this class hardly referred to other aspects in a decision statement than argumentation and organization. Class 3 was source-focused, assessors within this class were more likely to focus on source use and references. Class 4 was broadly focused, these assessors considered a great number of aspects of text quality when comparing texts.

The types of assessors are in line with research using absolute scoring procedures, where content and organization were mostly considered when assessing text quality (Vaughan, 1991; Huot, 1993; Sakyi, 2003; Wolfe, 2006). The language-focused class was related to the classes distinguished by Diederich et al. (1961) and Eckes (2008). Moreover, the source-focused class underpins Weigle and Montee's (2012) result that only some assessors consider the use of sources when assessing text quality. However, we did not find the same types of assessors as the other studies. This raises the question whether the method (comparative judgment) or the type of writing task impacted the determined types of assessors. For instance, in contrast to this study, the tasks used by Diederich et al. (1961); Eckes (2008), and Schaefer (2016) did not require the use of sources. This could explain the fact that a source-focused class was only found in our study, but not in other studies on rater types. Studies on how assessors adjust their focus according to the task they assess will improve our understanding of the stability of the types of assessors across tasks.

It is important to note that assessors were instructed to assess the full construct of text quality (argumentation, organization, language use, source use, language conventions, references, and layout). This study showed that for the validity of text scores, multiple assessors should be involved in a comparative judgment assessment. This increases the probability that the multidimensionality of text quality is represented in the text scores. To illustrate, the narrowly focused class rarely chose a text due to the quality of the source use, whereas the broadly- and source-focused class did. The latter group, however, rarely chose a text because of the language aspects. Combining the judgments of different types of assessors into scores leads to more informed text scores, representing the full complexity of text quality. Reading all these aspects of text quality in the decision statement underpins the argument that the final rank-orders represent the full construct of text quality in these assessments. In other assessments it might be of less importance that all of these aspects are considered. This depends on the aim of the assessment. For example, in some cases an assessment does not aim to assess students on the extent they apply language conventions. Most important for the validity argument is that students and assessors are aware of the assessment aim, so the aspects that are assessed by assessors are aligned with the student assignment.

Interestingly, the explanatory power of the types on the aspects that were assessed was rather limited. For example, all classes referred mostly to argumentation and organization and the four classes explained only for 2% whether argumentation was referred to in a decision statement (11% for layout). Also, none of the classes referred to one of the aspects of text quality in each and every decision statement, for example, the broadly focused class referred to organization in 71% of the decision statements. That means that for none of the classes one of the aspects of text quality was always the reason to choose for one text over the other. This conclusion seems to underpin the hypothesis that by offering assessors a comparison text, they rely less on their internalized ideal text but that their judgment is affected by the specific texts they are comparing. It makes us aware that more research is needed to establish what factors are at play. Pollitt and Murray (1996), Bartholomew et al. (2018), and Humphry and Heldsinger (2019) suggested that the quality of student works is related to the aspects upon which assessors focus. Comparing lower quality performance, lower order aspects as grammar and sentence structure seem to be more salient to assessors, whereas when comparing higher quality performance, the stylistic devices and audience are. Future research should consider both the assessor and characteristics of the text pair when looking into what aspects inform the comparison.

More research is also needed to better understand the implications for the resulting rank order of texts. Do assessors who belong to the same type make the same decisions on which texts are better? And does this differ from assessors belonging to another type? Studies on holistic and analytic scoring methods showed the relationship between what aspects were considered and resulting text scores (e.g., Eckes, 2012). But it is unknown whether this link can also be established within the context of comparative judgment. Unfortunately, the current data collection does not provide sufficient data to calculate text scores per assessor class.

The decision statements were shown to be a rich data source, enabling the detection of systematic differences between assessors. They were gathered during the assessment and did not interfere with the judgment process to a great extent. However, they only provided insight into the aspects that assessors revealed they based their judgments on, the just-noticeable-difference. They did not reveal information on the judgment process. To triangulate and extend the conclusions of this study, other data sources are required. Specifically, using think-aloud protocols while assessors make the comparisons would help us to understand what aspects assessors focus on when reading the texts, and how this relates to the aspects they subsequently base their decision on (Cumming et al., 2002; Barkaoui, 2011). This would enable us to gain insight into whether the narrowly focused and broadly focused classes also take other processing actions to reach a judgment. For example, Vaughan (1991) and Sakyi (2000) found that some assessors take only one or two aspects into account before deciding using an absolute holistic scoring procedure. Within the context of comparative judgment, this way of making decisions seemed to be typical for the whole narrowly focused class. Additionally, the broadly focused class, on average, referred to more aspects of text quality in a decision statement. This result suggests that these assessors apply a more analytical approach when comparing texts. Research into whether these differences in decision statements really reflect different processing strategies is needed to design comparative judgment in such a manner that it would optimally support assessors to make valid judgments.

CONCLUSION

We have concluded that different types of assessors can be distinguished based on differences in the aspects that the assessors were more likely to base their judgment on when comparing texts. These types have, however, only a small explanatory power regarding what aspects are assessed and all assessor' types had their main focus on organization and argumentation.

Nevertheless, the fact that we could detect assessor types implies that texts are ideally compared by multiple assessors—with different perspectives on text quality. Moreover, comparative judgment has been shown to be a promising way to integrate the judgments of multiple assessors into valid and reliable scores of text quality.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article are made available by the authors, without undue reservation through OSF (Lesterhuis et al., 2022).

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

ML and SD were responsible for the data collection. ML, RB, VD, and SD were responsible for the research design and conceptualization of the research questions. ML was responsible for the analyses. ML, RB, and TD were responsible for drafting the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Flanders Innovation and Entrepreneurship and the Research Foundation under (Grant No. 130043).

REFERENCES

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: an empirical study of their veridicality and reactivity. *Lang. Test.* 28, 51–75. doi: 10.1177/0265532210376379

Bartholomew, S. R., Nadelson, L. S., Goodridge, W. H., and Reeve, E. M. (2018). Adaptive comparative judgment as a tool for assessing open-ended design problems and model eliciting activities. *Educ. Assess.* 23, 85–101. doi: 10.1080/10627197.2018.1444986

Bejar, I. I. (2012). Rater cognition: implications for validity. *Educ. Meas. Issues Pract.* 31, 2–9. doi: 10.1111/j.1745-3992.2012.00238.x

Bradley, R. A., and Terry, M. E. (1952). Rank analysis of incomplete block designs the method of paired comparisons. *Biometrika* 39, 324–345. doi: 10.1093/biomet/39.3-4.324

Collins, L. M., and Lanza, S. T. (2009). *Latent Class And Latent Transition Analysis: With Applications In The Social, Behavioral, And Health Sciences*, Vol. 718. Hoboken, NJ: John Wiley & Sons.

Cumming, A., Kantor, R., and Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: a descriptive framework. *Mod. Lang. J.* 86, 67–96. doi: 10.1111/1540-4781.00137

Diederich, P. B., French, J. W., and Carlton, S. T. (1961). Factors in judgments of writing ability. *ETS Res. Bull. Ser.* 1961:98.

Eckes, T. (2008). Rater types in writing performance assessments: a classification approach to rater variability. *Lang. Test.* 25, 155–185. doi: 10.1177/0265532207086780

Eckes, T. (2012). Operational rater types in writing assessment: linking rater cognition to rater behavior. *Lang. Assess. Q.* 9, 270–292. doi: 10.1080/15434303.2011.649381

Heldsinger, S., and Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *Aust. Educ. Res.* 37, 1–19. doi: 10.1007/bf03216919

Humphry, S., and Heldsinger, S. (2019). Raters' perceptions of assessment criteria relevance. *Assess. Writ.* 41, 1–13. doi: 10.1016/j.asw.2019.04.002

Huot, B. (1990). Reliability, validity, and holistic scoring: what we know and what we need to know. *Coll. Composit. Commun.* 41, 201–213. doi: 10.1097/00001888-199404000-00017

Huot, B. A. (1993). "The influence of holistic scoring procedures on reading and rating student essays," in *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations*, eds M. M. Williamson and B. A. Huot (Creskhill, NJ: Hampton Press, Inc), 206236.

Jones, I., and Inglis, M. (2015). The problem of assessing problem solving: can comparative judgement help? *Educ. Stud. Math.* 89, 337–355. doi: 10.1007/s10649-015-9607-1

Jones, I., Swan, M., and Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *Int. J. Sci. Math. Educ.* 13, 151–177. doi: 10.1007/s10763-013-9497-6

Lesterhuis, M., Bouwer, R., van Daal, T., Donche, V., and De Maeyer, S. (2022). *Validity of Comparative Judgment Scores [dataset]*. OSF. doi: 10.17605/OSF.IO/8X692

Lesterhuis, M., van Daal, T., Van Gasse, R., Coertjens, L., Donche, V., and De Maeyer, S. (2019). When teachers compare argumentative texts: decisions informed by multiple complex aspects of text quality. *L1 Educ. Stud. Lang. Lit.* 18:1. doi: 10.17239/L1ESLL-2018.18.01.02

Luce, R. D. (1959). On the possible psychophysical laws. *Psychol. Rev.* 66, 81–95.

Messick, S. (1989). Meaning and values in test validation: the science and ethics of assessment. *Educ. Res.* 18, 5–11. doi: 10.3102/0013189x018002005

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assess. Educ. Princ. Policy Pract.* 19, 281–300. doi: 10.1080/0969594x.2012.665354

Pollitt, A., and Crisp, V. (2004). "Could Comparative Judgements Of Script Quality Replace Traditional Marking And Improve The Validity Of Exam Questions?," in *Proceedings of the British Educational Research Association Annual Conference, UMIST, Manchester, September 2004* (Cambridge: UCLES).

Pollitt, A., and Murray, N. L. (1996). "What raters really pay attention to," in *Performance Testing, Cognition and Assessment*, eds M. Milanovic and N. Saville (Cambridge: University of Cambridge), 7491.

Pollitt, A., and Whitehouse, C. (2012). *Using Adaptive Comparative Judgement To Obtain A Highly Reliable Rank Order In Summative Assessment*. Manchester: AQA Centre for Education Research and Policy.

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assess. Eval. High. Educ.* 34, 159–179. doi: 10.1080/02602930801956059

Sakyi, A. A. (2000). "Validation of holistic scoring for ESL writing assessment: How raters evaluate," in *Fairness And Validation In Language Assessment: Selected Papers From The 19th Language Testing Research Colloquium, Orlando, Florida* ed. A. J. Kunnan (Cambridge: Cambridge University Press), 129.

Sakyi, A. A. (2003). *Validation of Holistic Scoring for ESL Writing Assessment: How Raters Evaluate Compositions*. Ph.D. thesis. Toronto, ON: University of Toronto.

Schaefer, E. (2016). "Identifying rater types among native english-speaking raters of english essays written by japanese university students," in *Trends in Language Assessment Research and Practice: The View from the Middle East and the Pacific Rim* eds V. Aryadoust, and J. Fox (Cambridge: Cambridge Scholars), 184.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Pract. Assess. Res. Eval.* 9:4.

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., and De Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assess. Educ. Princ. Policy Pract.* 26, 59–74. doi: 10.1080/0969594x.2016.1253542

van Daal, T., Lesterhuis, M., Coertjens, L., van de Kamp, M. T., Donche, V., and De Maeyer, S. (2017). The complexity of assessing student work using comparative judgment: the moderating role of decision accuracy. *Front. Educ.* 2:44. doi: 10.3389/feduc.2017.00044

van den Bergh, M., Schmittmann, V. D., and Vermunt, J. K. (2017). Building latent class trees, with an application to a study of social capital. *Methodology* 13, 13–22. doi: 10.1027/1614-2241/a000128

van Weijen, D. (2009). *Writing processes, Text Quality, And Task Effects: Empirical Studies In First And Second Language Writing*. Dissertation, Netherlands Graduate School of Linguistics, Amsterdam.

Vaughan, C. (1991). "Holistic assessment: What goes on in the rater's mind," in *Assessing Second Language Writing In Academic Contexts* ed. L. Hamp-Lyons (Norwood, NJ: Ablex), 111–125.

Verhavert, S., Bouwer, R., Donche, V., and De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assess. Educ. Princ. Policy Pract.* 26, 541–562. doi: 10.1080/0969594X.2019.1602027

Verhavert, S., De Maeyer, S., Donche, V., and Coertjens, L. (2018). Scale separation reliability: what does it mean in the context of comparative judgment? *Appl. Psychol. Measur.* 42, 428–445. doi: 10.1177/0146621217748321

Vermunt, J. K., and Magidson, J. (2003). Latent class models for classification. *Comput. Stat. Data Anal.* 41, 531–537.

Wang, J., Engelhard, G., Raczynski, K., Song, T., and Wolfe, E. W. (2017). Evaluating rater accuracy and perception forintegrated writing assessments using a mixed-methods approach. *Assess. Writ.* 33, 36–47. doi: 10.1016/j.asw.2017.03.003

Weigle, S. C., and Montee, M. (2012). "Raters perceptions of textual borrowing in integrated writings tasks," in *Measuring Writing: Recent insights into Theory, Methodology and Practice*, eds E. VanSteendam, M. Tillema, G. C. W. Rijlaarsdam, and H. van den Bergh (Leiden: Koninklijke BrillNV), 117145.

Whitehouse, C. (2012). *Testing The Validity Of Judgements About Geography Essays Using The Adaptive Comparative Judgement Method*. Manchester: AQA Centre for Education Research and Policy.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometricscoring system. *Assess. Writ.* 4, 83–106. doi: 10.1016/S1075-2935(97)80006-2

Wolfe, E. W. (2006). Uncovering raters cognitive processing and focus using think-aloud protocols. *J. Writ. Assess.* 2, 37–56.

Conflict of Interest: ML and SD were co-founders of Comproved. However, this company was only founded after the research was conducted, analyses executed and first draft written (as a chapter in a dissertation).

The remaining 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Education

ORIGINAL RESEARCH
published: 04 July 2022
doi: 10.3389/feduc.2022.854378

Text Mining to Alleviate the Cold-Start Problem of Adaptive Comparative Judgments

Michiel De Vrindt[1]*, *Wim Van den Noortgate*[1,2] *and Dries Debeer*[1,2,3]

[1] Imec Research Group ITEC, KU Leuven, Kortrijk, Belgium, [2] Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium, [3] Faculty of Psychology and Educational Sciences, Ghent University, Ghent, Belgium

Comparative judgments permit the assessment of open-ended student works by constructing a latent quality scale through repeated pairwise comparisons (i.e., which works "win" or "lose"). Adaptive comparative judgments speed up the judgment process by maximizing the Fisher information of the next comparison. However, at the start of a judgment process, such an adaptive algorithm will not perform well. In order to reliably approximate the Fisher Information of possible pairs well, multiple comparisons are needed. In addition, adaptive comparative judgments have been shown to inflate the scale separation coefficient, which is a reliability estimator for the quality estimates. Current methods to solve the inflation issue increase the number of required comparisons. The goal of this study is to alleviate the cold-start problem of adaptive comparative judgments for essays or other textual assignments, but also to minimize the bias of the scale separation coefficient. By using text-mining techniques, which can be performed before the first judgment, essays can be adaptively compared from the start. More specifically, we propose a selection rule that is based both on a high (1) cosine similarity of the vector representations and (2) Fisher Information of essay pairs. At the start of the judgment process, the cosine similarity has the highest weight in the selection rule. With more judgments, this weight decreases progressively, whereas the weight of the Fisher Information increases. Using simulated data, the proposed strategy is compared with existing approaches. The results indicate that the proposed selection rule can mitigate both the cold-start. That is, fewer judgments are needed to obtain accurate and reliable quality estimates. In addition, the selection rule was found to reduce the inflation of the scale separation reliability.

Keywords: text mining, natural language processing, comparative judgments, educational assessment, computational linguistics, psychometrics, educational technology

OPEN ACCESS

Edited by:
Sven De Maeyer,
University of Antwerp, Belgium

Reviewed by:
Jinnie Shin,
University of Florida, United States
Elise Crompvoets,
Tilburg University, Netherlands

****Correspondence:***
Michiel De Vrindt
michiel.de.vrindt@gmail.com

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *13 January 2022*
Accepted: *30 May 2022*
Published: *04 July 2022*

Citation:
De Vrindt M, Van den Noortgate W and Debeer D (2022) Text Mining to Alleviate the Cold-Start Problem of Adaptive Comparative Judgments. Front. Educ. 7:854378. doi: 10.3389/feduc.2022.854378

1. INTRODUCTION

For rubric marking of students' works, assessors are required to isolate and accurately evaluate the criteria of the works. Grades or marks follow from how well certain criteria or the so-called 'grade-descriptors' are satisfied (Pollitt, 2004). Especially when the students' works are open-ended (e.g., essay text, portfolios, and mathematical proofs), rubric marking can be a difficult task for assessors (Jones and Inglis, 2015; Jones et al., 2019). Even when assessors are well-experienced, their assessments are likely to be influenced by earlier assessments, inevitably making the given grades relative to some extent. The method of comparative judgments (CJ), as introduced by Thurstone (1927), directly exploits the relative aspect of assessing open ended works. In CJ, rather than assessing individual works, pairs of works are holistically and repeatedly compared. That is, assessors (or judges) are not required to assign a grade on a specific (or multiple) grading scale(s); they only need to select the better work (i.e., the winner) of each pair that was assigned to them. Consequently, differences in rater severity (i.e., assessors that systemically score more severe or more lenient) and differences in perceived qualities between assessors become negligible (Pollitt, 2012). Based on the win-lose judgments of the comparisons, quality estimates of the students' works are obtained. As such, CJ allows a reliable and valid assessment of open-ended works that require subjective judgments. In addition to the capability of creating a valid and reliable quality scale, the process of CJ has proven to decrease the cognitive load that is required for the assessment process and develops the assessor's assessment skills (Coenen et al., 2018). From the students' perspective, CJ can include quantitative and qualitative feedback. Quantitative feedback is directly available from the final rank-order of essays, whereas quantitative feedback can be incorporated by including assessors' remarks (e.g., strong and weak points of essays). Hence, CJ can be used for both summative and formative assessments.

The original CJ algorithm pairs students' works randomly. A drawback of random pairings is that it typically requires many comparisons to obtain sufficiently reliable quality estimates. Consequently, the assessors' workload can be high. Several strategies have been proposed to minimize the number of pairwise comparisons while maintaining the reliability of the quality estimates and the final ranking of the works. Generally, these strategies try to make the repeated selection of pairs as optimal as possible (Rangel-Smith and Lynch, 2018; Bramley and Vitello, 2019; Crompvoets et al., 2020). For instance, Pollitt (2012) proposed a selection rule that speeds up the "scale-building" process by repeatedly selecting the pair for which a comparison would add the most information to the estimated qualities. More specifically, pairs are selected so that the expected Fisher Information of each next comparison is maximized based on the current quality estimates (refer to below). Because the quality estimates are repeatedly updated during the process, and because, based on the updated estimates, the most informative pair is repeatedly selected, this selection algorithm will be referred to as "adaptive comparative judgments" (ACJ).

Adaptive comparative judgments has two important shortcomings. First, at the start of the judgment process, the pairings cannot be made adaptively because quality estimates are only available after a minimal number of comparisons. This issue is typically referred to as the "cold-start problem." Current implementations of ACJ generally select the initial pairs randomly, where the adaptive selection starts only after these initial random pairings. Yet the first adaptive pairings are highly determined by the outcomes of the initial comparisons and judgments. Thus, if by chance low-quality works are paired with other low-quality works, it is possible that a low-quality work "wins" multiple initial comparisons, resulting in a high first quality estimate. When a low-quality work with a high first quality estimate is subsequently paired with a high-quality work (which is likely in the first ACJ-based pairings), the obtained judgment will have a limited contribution to the final quality estimate and ranking. Moreover, it may take multiple additional comparisons before the quality estimate of the low-quality work is properly adjusted and ACJ can have its beneficial impact. To prevent this behavior, Crompvoets et al. (2020) proposed a selection rule that introduces randomness in the selection of initial pairs while Rangel-Smith and Lynch (2018) selected initial pairs with more different initial quality estimates. Yet, although these selection rules may reduce the probability of strong distortions in ACJ, it also reduces the efficiency of the judgment process.

Second, the adaptive selection of pairs based on the maximum Fisher Information typically pairs work with similar true qualities. Therefore, low-quality works are often compared with other low-quality works and high-quality works with other high-quality works. These adaptive comparisons not only increase the reliability of the quality estimates (i.e., they lower the standard errors), but they also tend to inflate the estimated quality scale when the number of comparisons is still small (i.e., the estimated qualities are more extreme than the true qualities) (Crompvoets et al., 2020). The combination of lower standard errors and an inflated latent scale can cause inflation of the scale separation reliability (SSR), which is a commonly used estimator for the reliability of the obtained quality estimates (Bramley, 2015; Rangel-Smith and Lynch, 2018; Bramley and Vitello, 2019; Crompvoets et al., 2020). This is problematic because the SSR is typically used to decide when to stop the ACJ process. That is, the ACJ process is typically stopped when predefined reliability, as estimated by the SSR, is reached. When the SSR is overestimated due to the adaptive selection algorithm, there is a risk that the ACJ process is stopped prematurely. Indeed, Bramley (2015) reported that for true reliability of 0.70, an SSR of 0.95 may be expected when using ACJ. Moreover, Bramley and Vitello (2019) compared the quality estimates of the works that were obtained using ACJ and a limited number of comparisons per work, with quality estimates obtained by comparing every work to every other work ("all-by-all" design). The SD of the ACJ-obtained scale was 0.391 times larger than the SD of the all-by-all-obtained scale.

The issue of the SSR inflation in ACJ is widely known and some solutions have been proposed. These solutions consist of modifying the assessment design in order to increase the number of comparisons, add randomness to the adaptive selection algorithm or impose a minimal difference between quality parameter estimates to be selected (Rangel-Smith and Lynch, 2018; Bramley and Vitello, 2019; Crompvoets et al., 2021). Yet all strategies decrease the efficiency of the judgment process (i.e., more comparisons are required). Therefore, in this study, we explore a new strategy to alleviate the cold-start problem and reduce the SSR inflation in ACJ. We focus on the application of ACJ to assess textual works and propose the use of text-mining techniques to obtain numerical representations of the texts that capture semantic and syntactical information. Based on these numerical representations, the semantic and syntactical similarities of the texts can be computed. Because both the text mining techniques and the computation of the similarities can be performed before the start of the ACJ process, the initial pairings can be based on the similarities of the texts, rather than randomly pairing texts. As such, the cold-start problem and the SSR inflation may be mitigated. We explore different text mining techniques and evaluate our strategy using two sets of textual works.

In the remainder of this article, we first introduce the Bradley-Terry-Luce model (Bradley and Terry, 1952) for comparative judgment data and discuss the ACJ process in more detail (Pollitt, 2012). After presenting the SSR reliability estimator, the proposed text-mining strategy is explained, including the necessary text-pre-processing for extracting textual information. Different representation techniques are considered: term frequency-inverse document frequency (Aizawa, 2003), averaged word embeddings (Mikolov et al., 2013), and document embeddings (Le and Mikolov, 2014). Subsequently, we explain how the textual representations can be used to select initial pairs of essays by computing the similarity between the texts. More specifically, we propose a new progressive selection rule, in which the adaptive selection rule gradually becomes more important. We illustrate the proposed strategy using two real essay sets. Moreover, using simulated data the performance of the new progressive selection rule and the different text representation techniques is evaluated. The impact on the SSR inflation and the precision of the quality estimates is compared across conditions. After discussing the results, limitations and future research opportunities are discussed.

2. METHODS

2.1. Comparative Judgements-Design

2.1.1. Bradley-Terry-Luce Model

Let there be a set S of N works that should be assessed. Consider work i and work j with j and i in S. According to the Bradley-Terry-Luce model (BTL), the probability that work i wins over work j in a comparison, $\Pr(x_{ij} = 1)$, depends on the quality parameters θ_i and θ_j of work i and j, respectively (Bradley and Terry, 1952):

$$\Pr(x_{ij} = 1|\theta_i, \theta_j) = \frac{exp(\theta_i - \theta_j)}{1 + exp(\theta_i - \theta_j)}, \tag{1}$$

$$\text{where } x_{ij} \sim Bern(Pr(x_{ij} = 1)). \tag{2}$$

Based on the win-lose (i.e., 0, 1) data of many comparisons, the vector of all quality parameters $\boldsymbol{\theta}_{1\times N}$ can then be estimated by applying maximum-likelihood based methods to the BTL (Hunter, 2004).

2.1.2. Adaptive Comparative Judgement

When $\theta_i = \theta_j$ (i.e., the works i and j have equal quality parameters), then following Equation (1), the probability that work i wins over work j in a comparison is equal to $\Pr(x_{ij} = 1|\theta_i, \theta_j) = 0.5$. Moreover, the outcome for comparisons with $\Pr(x_{ij} = 1|\theta_i, \theta_j) = 0.5$ has the highest possible variance $\sigma^2(x_{ij} = 1) = \sigma^2(x_{ji} = 1) = 0.25$, and the expected Fisher information will be maximal. Therefore, the outcome of such a comparison will add the maximal amount of information to the estimation for the quality parameters (Pollitt, 2004). For ACJ as in Pollitt (2012), the works with the smallest difference in estimated quality parameters will be paired together, because the computed Fisher information is highest for these pairs.

Although the BTL allows multiple comparisons between pairs of works, CJ and ACJ typically restrict the number of comparisons per pair (by a single rater) to be maximally one: $x_{ij} = \{0, 1\}$ $(i \neq j)$. For N works, there are $\frac{N\times(N-1)}{2}$ unique comparisons. We denote this set of unique comparisons as B. In addition, let B_m be the set of unique pairs that is not yet compared after the mth judgment. Hence, generally in ACJ, the pair that will be selected for the $m+1$th comparison is the pair with the highest expected Fisher information $I(\hat{\theta}_i^{(m)}, \hat{\theta}_j^{(m)})$ (i.e., with the smallest distance between the quality estimates $\hat{\theta}_i^{(m)}$ and $\hat{\theta}_j^{(m)}$) in B_m.

Which pair has the highest Fisher information changes through the ACJ process because the quality estimates are continuously updated. Originally, Pollitt (2012) proposed to update all quality estimates $\hat{\boldsymbol{\theta}}_{1\times N}$ simultaneously after 'a round of comparisons'in which all works were compared once. However, because updating and re-estimating $\hat{\boldsymbol{\theta}}$ only after a certain number of comparisons results in a selection of pairs that are not based on the most up-to-date quality estimates (Crompvoets et al., 2020), $\hat{\boldsymbol{\theta}}$ is updated after every single comparison m in this study.

To repeatedly estimate the quality parameters after each comparison m, an expectation maximization algorithm is used (Hunter, 2004). Formally, for comparison $m + 1$ all qualities $\hat{\theta}_i \in \hat{\boldsymbol{\theta}}$ for work $i, \ldots, N$ are estimated using:

$$\hat{\theta}_i^{(m+1)} = \log\left(x_i \left(\sum_{j\neq i}^{N} \frac{n_{ij}}{e^{\hat{\theta}_i^{(m)}} + e^{\hat{\theta}_j^{(m)}}}\right)^{-1}\right) \tag{3}$$

$$\hat{\theta}_i^{(m+1)} = \hat{\theta}_i^{(m+1)} - \frac{\sum_i^N \hat{\theta}_i^{(m)}}{N} \tag{4}$$

where n_{ij} is an indicator variable indicating whether work i and j are compared yet and x_i is the total number of wins of work i.

After updating every $\hat{\theta}_i^{(m+1)}$, all quality parameters are centered so that the mean of the quality estimates will be zero (Equation 4). If the work has not been compared yet or it loses every comparison, its quality estimate is unidentifiable. To make the quality parameters identifiable, a small quantity is added to x_{ij} (i.e., 10^{-3}) (Crompvoets et al., 2020).

2.1.3. Stochastic Adaptive Comparative Judgments

In the original ACJ algorithm by Pollitt (2012), only the point estimates of the quality parameters are considered in the selection algorithm. However, the uncertainty of these point estimates can be large, especially at the beginning of the ACJ process when there are few judgments per work. In order to also consider the uncertainty of the quality estimates, Crompvoets et al. (2020) included the standard error of the quality estimate in the selection algorithm. That is, for comparison $m + 1$, first the work i with the largest standard error of the quality estimate $\hat{\sigma}_{\hat{\theta}_i^{(m)}}$ is selected.

Then, rather than selecting the work j for which $I(\hat{\theta}_i^{(m)}, \hat{\theta}_j^{(m)})$ is maximized (with the comparison of i and j still in B_m), the work j is randomly selected from all candidates left in B_m with a probability that is a function of the distance between $\hat{\theta}_i^{(m)}$ and $\hat{\theta}_j^{(m)}$, and $\hat{\sigma}_{\hat{\theta}_i^{(m)}}$. More specifically, the selection probabilities are proportional to the densities of the $\hat{\theta}_j^{(m)}$ in a normal distribution with mean $\hat{\theta}_i^{(m)}$ and variance $\hat{\sigma}^2_{\hat{\theta}_i^{(m)}}$ (Crompvoets et al., 2020).

This adaptive selection rule is stochastic and introduces randomness to the algorithm. If few comparisons have been made with work i, the normal distribution of the quality parameter will have wider tails, which causes the selection rule to be more random. As more comparisons are made, the normal distribution will become more peaked and student works with similar quality parameters will be selected with a higher probability. A drawback of this algorithm is that only $\hat{\sigma}_{\hat{\theta}_i^{(m)}}$ is considered. $\hat{\sigma}_{\hat{\theta}_j^{(m)}}$ is not taken into account.

To compute the standard error of a quality parameter estimate $\hat{\sigma}^{(m)}_{\hat{\theta}_i}$ after each comparison, the observed Fisher Information function with respect to $\hat{\boldsymbol{\theta}}^{(m)}$ given all the judgment outcomes $\mathbf{x}$ is used:

$$\hat{\sigma}_{\hat{\theta}_i} = \left(-\frac{\partial^2 \ell(\hat{\boldsymbol{\theta}}|\mathbf{x})}{\partial \hat{\theta}_i^2}\right)^{-1/2} \tag{5}$$

$$= \sum_{j \neq i}^{N} \left(\frac{x_{ij}\, e^{\hat{\theta}_i - \hat{\theta}_j}}{(1+e^{\hat{\theta}_i - \hat{\theta}_j})^2} + \frac{x_{ji}\, e^{\hat{\theta}_j - \hat{\theta}_i}}{(1+e^{\hat{\theta}_j - \hat{\theta}_i})^2}\right)^{-1/2} \tag{6}$$

where x_{ij} is 1 when work i wins the comparison over j ($x_{ij} = 1 - x_{ji}$). In Equation (6), superscript (m) is dropped for the ease of reading. In this article, the 'stochastic ACJ' selection rule by Crompvoets et al. (2020) is used for all ACJ.

2.1.4. SSR as Reliability Estimator

If the true quality parameters $\boldsymbol{\theta}$ of a set of works are known, the reliability of the estimated qualities $\hat{\boldsymbol{\theta}}$ can be obtained from the squared Pearson correlation of the true quality and estimated parameters $\rho^2_{\boldsymbol{\theta},\hat{\boldsymbol{\theta}}}$. This corresponds to the ratio of the variance of the true quality levels and the variance of the estimated quality parameters. The more similar the variances are, the higher the reliability will be. In practice, the reliability of the assessment is an important criterion. Often, a minimum value for reliability is required. In real assessment situations, however, the true quality parameters are not available, which makes it impossible to compute the reliability as $\rho^2_{\boldsymbol{\theta},\hat{\boldsymbol{\theta}}}$.

An estimator for the reliability that can be computed without the true quality parameters is the Scale Separation Reliability (SSR), which is based on the estimated quality parameters and their uncertainty (Brennan, 2010). To compute the SSR, the unknown true variance of the quality parameters, denoted σ^2, is approximated by the difference between the variance of the quality estimates, denoted $\hat{\sigma}^2$, and the mean squared error of the standard errors of the quality estimates, $\hat{\sigma}_{\hat{\theta}_i}$. The SSR is defined as:

$$SSR = \frac{\hat{\sigma}^2 - \mathrm{MSE}(\hat{\sigma}_{\hat{\theta}_i})}{\hat{\sigma}^2} \tag{7}$$

$$\text{with } \mathrm{MSE}(\hat{\sigma}_{\hat{\theta}_i}) = \mathrm{E}(\hat{\sigma}^2_{\hat{\theta}_i}). \tag{8}$$

Equation (7) indicates that a higher variance of the quality estimates and smaller standard errors of the estimates will lead to a higher SSR. For the full derivation of the SSR, refer to Verhavert et al. (2018). For the SSR to be estimable, $\hat{\sigma} > 0$ and $\hat{\sigma} \geq \mathrm{E}(\hat{\sigma}_{\hat{\theta}_i})$ must hold.

2.1.5. Vector Representations of Essays

Numerical representations of texts should capture the most important features of the texts, both with respect to syntax and semantics. Statistical language modeling allows the mapping of natural unstructured text to a vector of numeric values. We consider three representation techniques to represent essay tests as numerical vectors: term frequency-inverse document frequency ("tf-idf") (Aizawa, 2003), averaged word embeddings (Mikolov et al., 2013), and document embeddings (Le and Mikolov, 2014). A brief explanation of the construction of the three representation techniques will be given.

First, tf-idf representations are constructed based on word frequencies: the relative frequency of words in a document is offset by how often words appear across documents (Aizawa, 2003). A word that occurs frequently in a document but that doesn't occur often in other documents, receives a higher weight. However, because it only considers word frequencies, tf-idf is limited in terms of extracting syntactical meanings. One way to extract syntactical information is by grouping sequences of words that often occur together, called "n-grams." Yet even in the case of n-grams, tf-idf representations do not incorporate the syntactical meaning of texts apart from relations between n-grams. In addition, because every word (or n-gram) across the documents corresponds to one dimension, tf-idf representations are typically highly dimensional.

Second, average word embeddings are a more complex representation technique that incorporates syntactical information and that is not highly dimensional. Average word embeddings are distributional representations based on the so-call "skip-gram word embeddings" neural network architecture. In the skip-gram architecture, a shallow neural network is constructed with a word as input and its surrounding words as output (Mikolov et al., 2013). The "embeddings" are the weights of the hidden layer in the neural network, which are obtained from predicting the set of surrounding words for each input word. The predicted surrounding words are the words that have the largest probability on average as given by the sigmoid function of the dot product of the embeddings of each surrounding word with the input word. However, iterating over all possible combinations of surrounding words and calculating probabilities is computationally intensive. As an alternative, the objective function is minimized by correctly distinguishing between surrounding words and sampled non-surrounding words (i.e., "negative sampling"). Ultimately, essay representations are obtained from the average pooling of the word embeddings of all the words in each essay. A disadvantage of averaged word embeddings is that it does not account for the dependence of the meaning of words coming from the document (or essay) they are part of.

Finally, document embeddings are an extension of word embeddings that allow for this document-dependence (Le and Mikolov, 2014). Instead of learning embeddings on the level of words and aggregating it to embeddings of documents, document embeddings can be learned directly. The distributed continuous-bag-of-words architecture are neural networks that predict whether words occur in a given document. The words are those with the highest probability on average as given by the sigmoid function of the dot product of a document embedding and word embeddings. Negative sampling is also possible by sampling words that do not occur in a given document. The distributed bag-of-words architecture for document embeddings can be initialized by a pre-trained set of word embeddings (Tulkens et al., 2016). The pre-trained model consists of embeddings of words that are trained on a very large corpus of texts. The reason for using a large corpus is that words can be learned from or 'embedded' in many different contexts. Pre-trained models are often used in natural language processing as sample corpora are often not large enough. If the contexts in which words are learned are very different from those in the essay texts, then the pre-trained word embeddings would not be fit. However, this possibility is only small as pre-trained corpora are very large.

The main differences between the three representations are three-folded. First, the dimensions of the vector representation can have either an explicit interpretation based on term frequencies (tf-idf) or an implicit interpretation (averaged word embeddings and document embeddings). Second, the length of the vector can be variable (tf-idf) or fixed (averaged word embeddings and document embeddings). Finally, the representations can be sparse with many zero dimensions (tf-idf) or dense with few zero dimensions (averaged word embeddings and document embeddings).

When comparing average word embeddings with document embeddings, document embeddings have a clear advantage, which is apparent from the clustering of the embeddings in vector space. Document embeddings tend to be located close to the embeddings of the keywords of the document (Lau and Baldwin, 2016). Average word embeddings, on the other hand, tend to be located at the centroid of the word embeddings of all the words in a document. However, document embeddings are not free of issues. Ai et al. (2016) pointed out that shorter documents can be overfitted and often show too much similarity; the sampling distribution used in the document embeddings is improper in that frequent words can be penalized too rigidly; and sometimes document embeddings do not detect synonyms of words in different documents even though the context is alike. Despite these issues, document embeddings showed better results for various tasks when compared to tf-idf or averaged word embeddings (Le and Mikolov, 2014). Therefore, we expect that use document embeddings to represent essays and select pairs of essays based on these representations to outperform the tf-id and average word embeddings.

2.2. Progressive Selection Rule Based on Vector Similarities

In this manuscript, we propose a progressive selection rule that combines the stochastic ACJ selection of Crompvoets et al. (2020) with a similarity component based on the cosine similarities of the vector representations of essays. Initially, the progressive selection rule selects pairs based on the similarity of their representations (i.e., how close they are to each other in vector space). As more judgment outcomes become available, the weight of the 'stochastic adaptivity' component increases so that pairs are increasingly selected based on the quality parameter estimates of the works.

To quantify the similarity between the vector representations, the cosine similarity is chosen over the Euclidean distance and Jaccard similarity. First, unlike the Euclidean distance, the cosine similarity is a normalized measure (with range $[-1, 1]$). Second, although also normalized, the Jaccard similarity tends to not work well for detecting similarities between texts when there are many overlapping words between essays (Singh and Singh, 2021). The cosine similarity between two works i and j is the cosine of the angle of their corresponding vector representations $\boldsymbol{\gamma}_i$ and $\boldsymbol{\gamma}_j$:

$$S(\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_j) = \frac{\boldsymbol{\gamma}_i \cdot \boldsymbol{\gamma}_j}{||\boldsymbol{\gamma}_i||||\boldsymbol{\gamma}_j||}. \tag{9}$$

Note that $\boldsymbol{\gamma}_i$ is of variable-length for tf-idf representations and fixed-length for averaged word embeddings and document embeddings. The fixed length is determined by the dimensionality of the pre-trained word embeddings which in this case is 320 (Tulkens et al., 2016).

For the similarity component in the progressive selection rule, the cosine similarities of all works j with respect to work i are non-linearly transformed so that higher similarities are up-weighted and lower similarities are down-weighted. This can be achieved by assigning the probability mass of the CDF of a normal distribution to all cosine similarity values of works j with respect to work i. To encourage the selection of pairs with very high similarities (which can be rare) an upper quantile of the cosine similarities is chosen as the mean of the normal CDF. The quantile will function as a (soft) threshold parameter. So the probability to select works j with any lower similarity value than the quantile will be close to 0. A second component is the stochastic adaptive selection rule as in Crompvoets et al. (2020) (refer to above). As such, the parameter uncertainty of work i can be taken into account for the selection of work j.

The cosine similarity also measures dissimilarities (i.e., negative values). However, dissimilarities are uninformative for the pairing of essays, and negative values cannot be used as probabilities in the progressive selection rule. Hence, the cosine similarities are truncated at 0.

The two components are combined in the progressive selection rule as follows: a pair $\{i, j\}$ is selected from B_m so that work i has the minimum number of comparisons out of all the works, and work j is sampled with a probability given by the weighted sum of the similarity and the adaptivity component. The weights depend on the number of times work i has been compared. Formally, at the $m_i + 1$-th comparison of work i it is paired with work j given the probability mass function:

$$\Pr(j|i) = \frac{(1-w_i)\,\Phi\left(S(\boldsymbol{\gamma}_i,\boldsymbol{\gamma}_j) - Q_{\boldsymbol{S}_i}(p)\right)}{\sum_{\{i,j\}\in B_m}\Phi\left(S(\boldsymbol{\gamma}_i,\boldsymbol{\gamma}_j) - Q_{\boldsymbol{S}_i}(p)\right)} + \frac{w_i\,\phi\left(\frac{\hat{\theta}_j-\hat{\theta}_i}{\hat{\sigma}_{\hat{\theta}_i}}\right)}{\sum_{\{i,j\}\in B_m}\phi\left(\frac{\hat{\theta}_j-\hat{\theta}_i}{\hat{\sigma}_{\hat{\theta}_i}}\right)} \tag{10}$$

where Φ is the CDF of a standard normal distribution with as mean the p-th quantile of all cosine similarities with the essay i except itself, $Q_{\boldsymbol{S}_i}(p)$ with $\boldsymbol{S}_i = (S(\boldsymbol{\gamma}_i,\boldsymbol{\gamma}_j),\ldots,S(\boldsymbol{\gamma}_i,\boldsymbol{\gamma}_{N-1}))$. For the adaptive component, the density values of all $\hat{\theta}_j$ for the normal distribution with mean $\hat{\theta}_i$ and standard error $\hat{\sigma}_{\hat{\theta}_i}$ are taken. The weight $w_i \in [0,1]$ of work i depends on m_i (this is the number of times work i has been compared) and on m_d (this is the minimal desired number of comparisons for each work) with $m_i \leq m_d$ and decay parameter t ($t > 0$) as follows:

$$w_i = \begin{cases} 0 & \text{if } m_i = 0, \\ (\frac{m_i}{m_d})^t & \text{otherwise.} \end{cases} \tag{11}$$

If $m_i = 0$, work i is compared for the first time and will be allocated only based on the similarity component. Moreover, one needs to determine the speed at which the weight of the similarity selection rule decays in favor of the adaptive component by setting the parameter t. In computerized adaptive testing, where progressive selection rules with a random component have been proposed, $t = 1$ is often chosen, which corresponds with a linear decrease of the weight of the random component (Revuelta and Ponsoda, 1998; Barrada et al., 2010). In this study, however, we tune the decay parameter t to find the optimal progressive rule. A higher t leads to a slower decrease in the similarity component, whereas a smaller t leads to a faster decrease of the similarity component. For $t = 0$, the progressive rule reduces to the stochastic ACJ selection rule.

2.3. Experiment

2.3.1. Datasets: Essay Sets

The proposed selection rule will be tested on two different essay sets. The essay sets along with quality scores were provided by the company Comproved. The qualities of these essays were estimated from CJ-assessments and are centered around zero. For this study, these are assumed to be the true quality levels, which is a reasonable assumption given that each essay was compared up to 20 times with random CJ. Both essay sets are of a similar size although the length of the essays in essay set 1 is more variable than those in essay set 2 (refer to **Table 1**). The quality levels show a symmetric distribution around zero. For essay set 1, 16-year-old students were asked to write a two-page research proposal on a topic of their choice. For essay set 2, 16-year-old students needed to write a two-page argumentative essay about the conservation of wildlife. For both essay sets, the true quality levels show only a small spread. This corresponds to assessment situations where it would be hard for the assessors to discriminate between the quality levels of essays (Rangel-Smith and Lynch, 2018).

TABLE 1 | Description of the contents of two essay sets.

	Essay set 1	Essay set 2
Assignment	Research proposal	Argumentation
N	141	150
SD of qualities	1.66	1.13
Range of qualities	−5.42, 4.92	−3.62, 2.10
Proportion qualities ≤ 0	0.49	0.45
Proportion qualities > 0	0.51	0.55
Total # of words	67340	58037
Avg. length essays	474	386

2.3.2. Preprocessing of Essay Texts

The initial preprocessing steps on the essay texts are common for every representation technique and they are in accordance with the steps performed on the pre-trained SoNaR corpus (Oostdijk et al., 2013; Tulkens et al., 2016). This involves lowercasing, removing punctuations, removing numbers, removing single letter words, and decoding utf-8 encoding. The only single letter word that is included is "u" which is a Dutch formal pronoun. In contrast to Tulkens et al. (2016), we chose to also include sentences shorter than 5 words. The reason being that the essay set is short (1 or 2 pages) so every sentence may be meaningful (**Table 1**).

Some additional preprocessing steps on the texts depend on the representation technique. For the tf-idf representation of the essays, the essay texts will be normalized to a higher extent. This is necessary as the size of the essay sets is relatively small and no pre-trained corpus can be used with tf-idf. Extended normalization will decrease the length of the vocabulary, and hence, increase the similarities between essays. However, there may be a loss of information as well. A first additional step is the lemmatization of the words so that they are simplified to their root word, which is an existing word—unlike with stemming. In addition, for tf-idf the syntactical structures will be represented to some extent by allowing bi-grams of word pairs that often occur together. Including n-grams also decreases the high dimensionality of the vector representation because the vocabulary size decreases. Note that for the tf-idf representations, the idf-term is smoothed in order to prevent zero division (Aizawa, 2003).

For the representation of essays based on averaged word embeddings and document embeddings, the pre-trained SoNaR corpus with embeddings of Dutch words is used (Tulkens et al., 2016). The pre-trained corpus consists of 28.1 million sentences and 398.2 million words from various media outlets (news stories, magazines, auto-cues, legal texts, Wikipedia, etc.) (Oostdijk et al., 2013). The embeddings were learned using a skip-gram architecture with negative sampling (Mikolov et al., 2013). The embeddings have 320 dimensions. The pre-trained SoNaR corpus showed excellent results for training word embeddings in Tulkens et al. (2016). Note that this pre-trained corpus only contains correctly spelled words. This implies that misspelled words in the essays will not be represented, which may decrease their usability for making pairs. Also, grammatical mistakes can have an influence on the essay embeddings because word embeddings and document embeddings are sensitive to word order as it used for their training (Mikolov et al., 2013; Le and Mikolov, 2014). Preprocessing techniques like lemmatization or stemming are not performed for these representations to keep the essays closest to their original semantical and syntactical meaning. This is feasible given that almost all words can be found in the large pre-trained SoNaR corpus (Oostdijk et al., 2013).

TABLE 2 | Quantiles of the cosine similarities between essays using different essay representation techniques for two essay sets.

Essay representation	Quantile (%)	Essay set 1	Essay set 2
Tf-idf	50	0.12	0.21
	70	0.14	0.23
	80	0.15	0.24
	90	0.17	0.26
Averaged word emb.	50	0.23	0.3
	70	0.30	0.37
	80	0.34	0.42
	90	0.40	0.51
Document emb.	50	0.22	0.22
	70	0.24	0.24
	80	0.27	0.26
	90	0.28	0.28

2.3.3. Baseline Selection Rules and Simulation Design

Three baseline selection rules will be tested: the random CJ, the stochastic ACJ as in Crompvoets et al. (2020), and a progressive selection rule with a random component for the initial comparisons. For the progressive rule with a similarity component (Equation 10), three essay representation techniques will be considered (i.e., tf-idf, averaged word embeddings, and document embeddings) and a progressive rule with a random component instead of a similarity component. The progressive selection rule with a random component is constructed to evaluate whether the similarity component in the progressive selection rule is more informative for the initial pairing of works than random pairs.

The performance of each selection rule will be assessed based on the SSR, the true reliability, and the SSR bias (their difference) for a given number of comparisons per work on average. Next, differences in SSR between the proposed progressive rule and the baseline selection rules will be evaluated based on the two

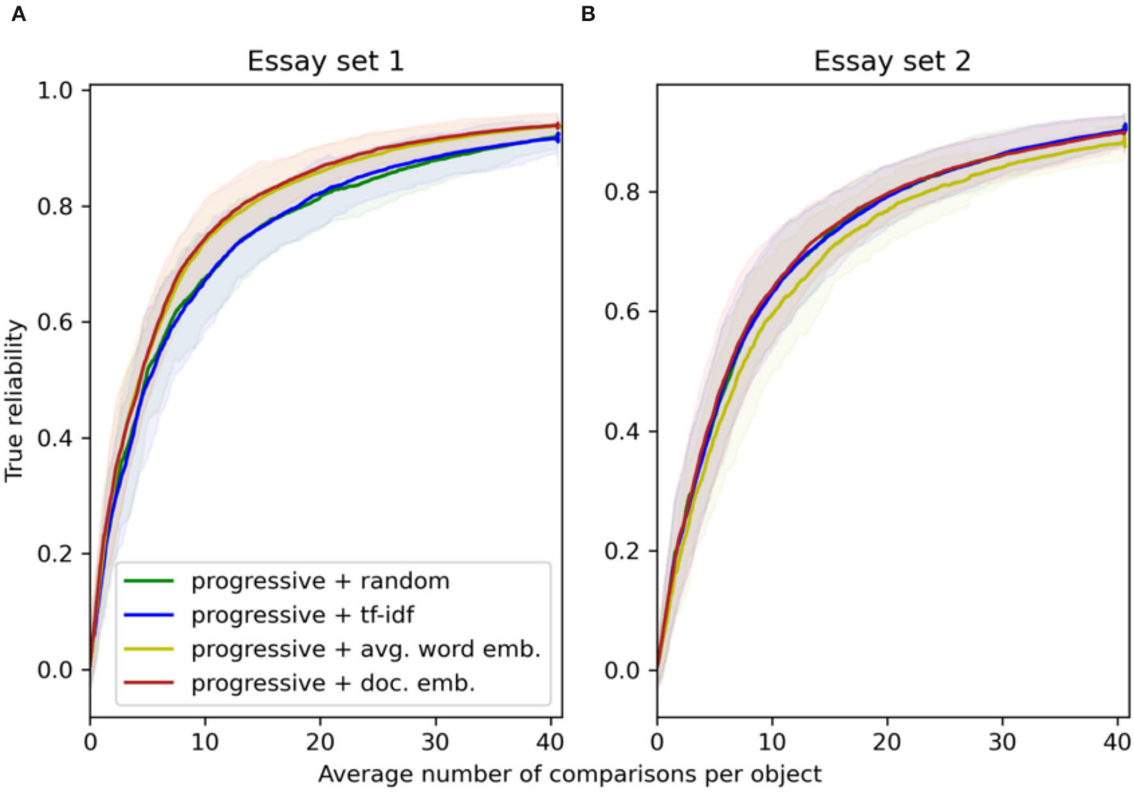


FIGURE 1 | For essay set 1 (**A**, $N_1 = 141$) and 2 (**B**, $N_2 = 150$), the true reliability resulting from the progressive selection with a random component, and a similarity component using tf-idf, averaged word embeddings and document embeddings (100 simulations). The solid lines indicate the mean values and the transparent bands indicate the 95% point-wise confidence intervals.

components that determine the SSR, namely the spread of the quality parameter estimates and their standard errors with respect to the ranking of essays (Equation 7). For brevity, not all representation techniques will be compared to the baseline selection rules here, only the one that performs the best in terms of SSR.

To simulate the judgment process the probability that work i wins as obtained from BTL-model (Equation 1) is compared to a random number drawn from a continuous uniform distribution between 0 and 1 (Davey et al., 1997; Crompvoets et al., 2020). If the probability is higher than the random value, work i wins the comparison. If the sampled value is smaller, work j wins the comparison. As such, one can imitate the stochastic process of judging. For each of the selection rules, the judgment process will be simulated 100 times (Matteucci and Veldkamp, 2013; Rangel-Smith and Lynch, 2018). A minimum of 40 work comparisons for all works is defined as a stopping rule ($m_d = 40$). This can show the asymptotic behavior of the SSR estimator for the different selection rules. For a minimum of 40 work comparisons per work, at least 50% of the possible pairings are compared given that $N_1 = 141$ and $N_2 = 150$. Preliminary simulations are conducted to tune the decay parameter t and the quantile p of the cosine similarities (Equation 10). That is, the true reliability and the SSR bias are evaluated for a grid of every parameter combinations for $p = \{0.50, 0.70, 0.80, 0.90, 0.95\}$ and $t = \{0.20, 0.40, 0.60, 0.80, 1.00, 2.00\}$. For each condition ($5 \times 5$) 50 simulations are conducted.

3. RESULTS

3.1. Tuning of the Decay Parameter and the Cosine Similarity Quantile

The preliminary simulations showed that a decay parameter (t) of 0.4 and a cosine similarity quantile (p) from 70 to 90% result in the highest SSR with a small bias (below 0.05). The 80% upper quantile of the cosine similarities was chosen. The cosine similarity corresponding to the 80% quantile is the smallest for tf-idf (0.15 and 0.24 for essay set 1 and 2, respectively) and the largest for averaged word embeddings (0.34 and 0.42 for essay set 1 and 2, respectively) (**Table 2**). The 80% quantile of the cosine similarities using document embeddings is 0.27 and 0.26 for essay set 1 and 2, respectively.

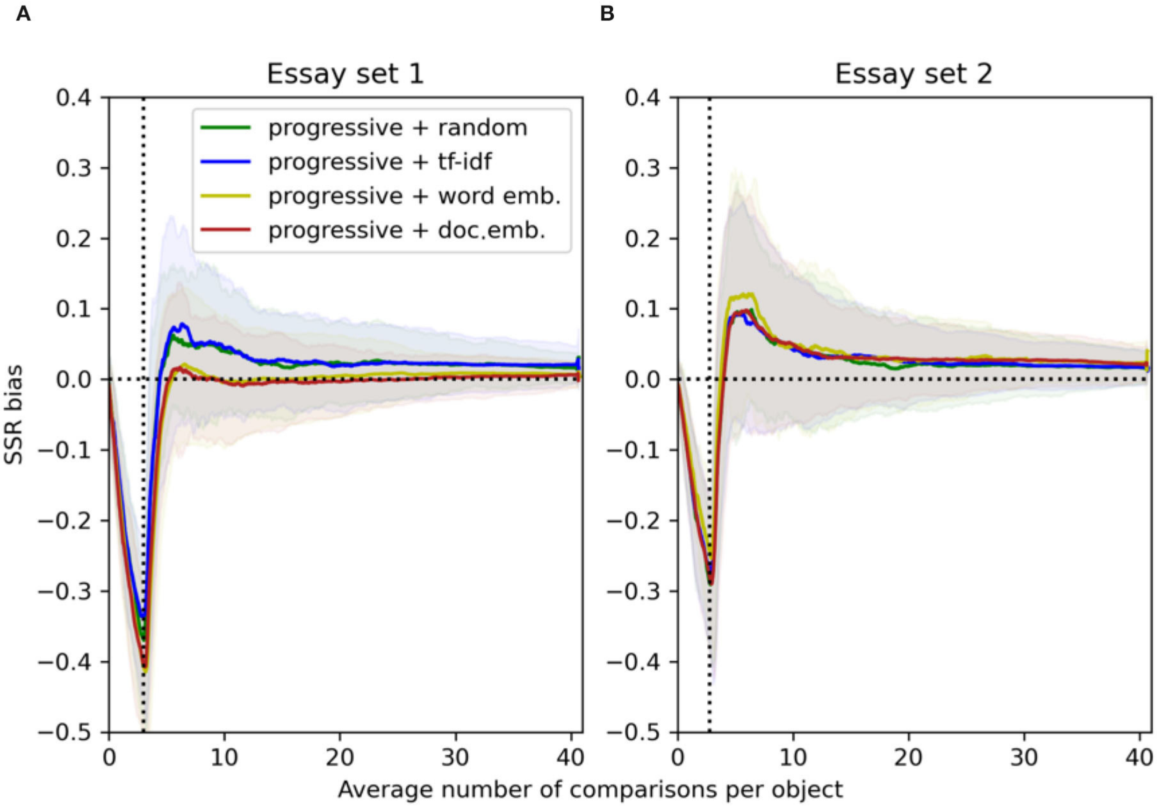


FIGURE 2 | For essay set 1 (**A**, $N_1 = 141$) and 2 (**B**, $N_2 = 150$), SSR bias resulting from the progressive selection with a random component, and a similarity component using tf-idf, averaged word embeddings and document embeddings (100 simulations). The solid lines indicate the mean values and the transparent bands indicate the 95% point-wise confidence intervals.

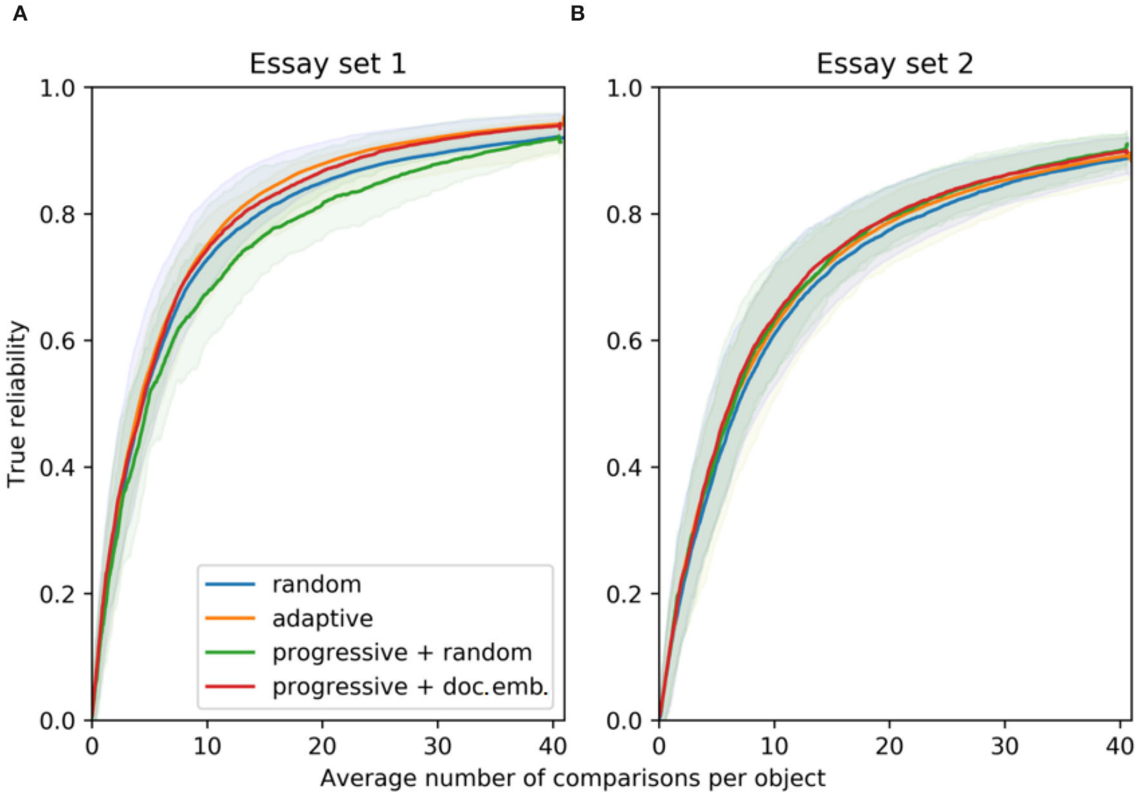


FIGURE 3 | For essay set 1 (**A**, $N_1 = 141$) and 2 (**B**, $N_2 = 150$), the true reliability resulting from the selection rules: for random (CJ), adaptive (ACJ), and a progressive selection with a random and a similarity component using document embeddings of essays (100 simulations). The solid lines indicate the mean values and the transparent bands indicate the 95% point-wise confidence intervals.

3.2. Performance of SSR Estimator

We will first describe the performance of the SSR estimator for the proposed progressive selection rule with different essay representation techniques. Subsequently, we will compare the progressive selection rule with the best performing representation technique to the CJ and ACJ baseline selection rules.

3.2.1. Performance of SSR for the Progressive Selection Rules

The performance of the SSR for the progressive rule with a similarity component is highly dependent on the chosen essay representation technique. For essay set 1, a similarity component based on averaged word embeddings and document embeddings seems to perform equally well in terms of reliability and SSR bias (**Figures 1A**, **2A**). The progressive selection rule with a similarity component based on tf-idf representations results in small true reliability similar to the progressive selection rule with a random component. This indicates that the similarities based on the tf-idf representations of essay set 1 are close to being random. However, for essay set 2 the progressive selection rule with a similarity component based on tf-idf performs better than with a random component, and unexpectedly, better than with a similarity component based on averaged word embeddings (**Figures 1B**, **2B**). For both essay sets, the progressive rule with a similarity component based on document embeddings performs at least as good as the progressive rule based on tf-idf or averaged word embeddings, and is always better than the progressive rule with a random component. This indicates that initial pairings based on the large cosine similarities of document embeddings can be beneficial.

The progressive rule with a similarity component produces higher true reliability than random CJ (**Figure 1**). When the similarity component is computed based on document embeddings, true reliability is reached that is 0.02–0.03 higher than for random CJ. The true reliability under the progressive rule with a similarity component is close to the high reliability under ACJ. Compared to ACJ, however, the progressive rule with a similarity component has an SSR bias that converges faster to below 0.05. For essay set 1, the SSR bias is even smaller than for random CJ (**Figure 2A**). For essay set 2, the SSR bias is more persistent than for random CJ which may be due to the smaller spread of its true quality levels (**Figure 2B** and **Table 1**).

3.2.2. Performance of SSR for the Baseline Selection Rules

The performance of the CJ and ACJ baseline selection rules in terms of the SSR is as expected given the average number of

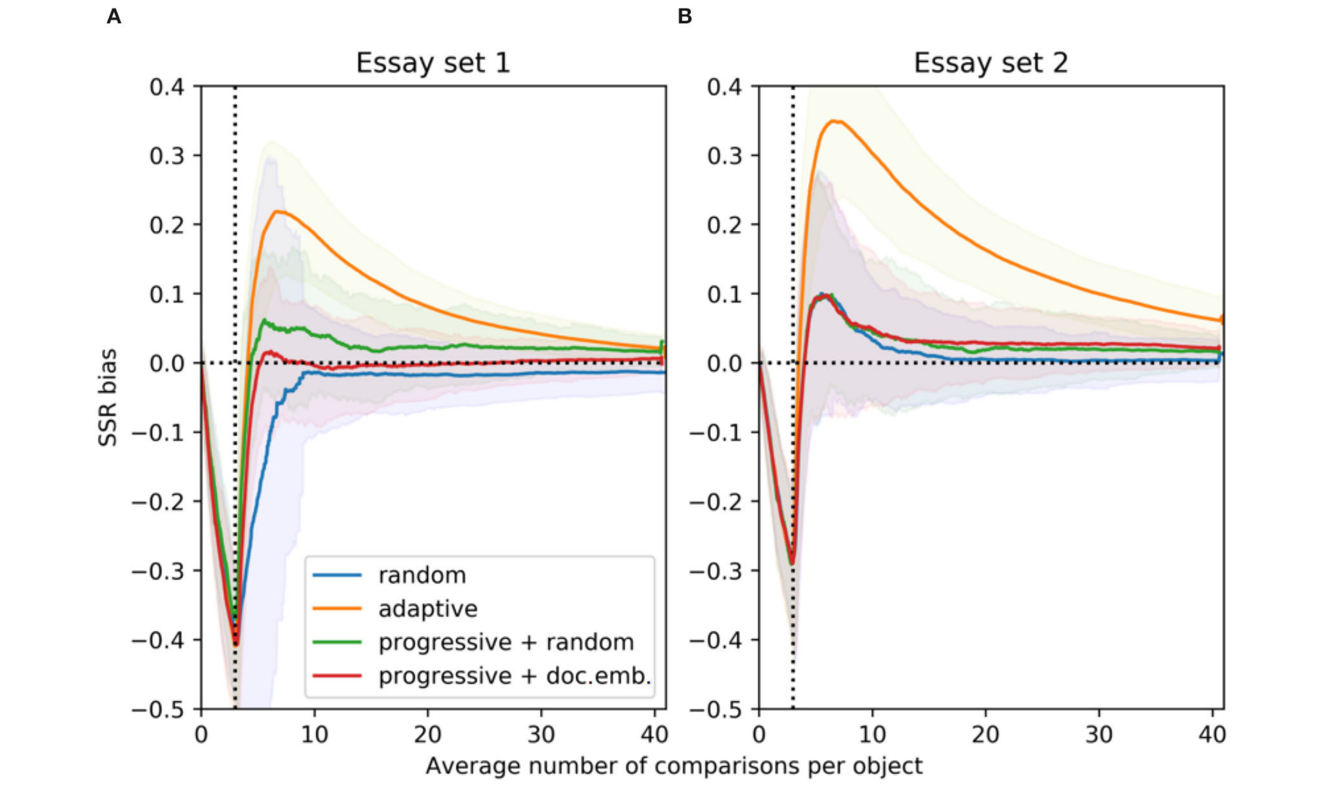


FIGURE 4 | For essay set 1 (**A**, $N_1 = 141$) and 2 (**B**, $N_2 = 150$), SSR bias resulting from the selection rules: random (CJ), adaptive (ACJ) and a progressive selection with a random and a similarity component using document embeddings (100 simulations). The solid lines indicate the mean values and the transparent bands indicate the 95% point-wise confidence intervals.

comparisons. The random CJ can result in an SSR that can both under- and over-estimate the true reliability at the start of the CJ process (**Figures 3**, **4**). Crompvoets et al. (2021) also reported positive SSR bias for the random CJ selection rule. The SSR bias for random CJ converges to <0.05 after on average 5 comparisons per work. In other words, up to 355 and 375 comparisons were needed for essay set 1 and 2, respectively. ACJ on the other hand results in an SSR that clearly overestimates the true reliability. After on average 10 comparisons per work, the SSR is 25% larger than the true reliability for essay set 1 (**Figure 3A**), and 52% for essay set 2 (**Figure 3B**). For ACJ, the SSR bias is only negligible (below 0.05) after on average 20 comparisons per work for both essay sets (**Figure 4**). Both baseline selection rules show evidence that their SSR is asymptotically unbiased—although the rate at which the bias reduces is the highest for random CJ. Note that for all selection rules, the SSR bias is negative until on average 5 comparisons per work are made. Even though ACJ produces inflated SSR estimates, it can produce true reliability that is 0.02–0.03 higher than for random CJ (**Figure 3**). This is already observed for more than 5 comparisons per work on average. The performance of the SSR for random CJ and ACJ is similar to in Crompvoets et al. (2020) and Rangel-Smith and Lynch (2018).

The results for the true reliability and the SSR produced by the progressive rule with a random component are inconsistent between essay sets. For essay set 1, the progressive rule with a random component results in quality parameter estimates that have the lowest true reliability out of all the selection rules (**Figure 3A**). For essay set 2, the progressive rule with a random component results in true reliability that is higher than for the random CJ and ACJ (**Figure 3B**). For both essay sets, the SSR bias for the progressive rule with a random component is smaller than for ACJ but larger than for random CJ (**Figure 4**).

The progressive selection rule with a similarity component based on document embeddings requires fewer judgments per work to reach the desired reliability (for instance, 0.70 or 0.80). For essay set 1, this progressive selection rule can reach reliability of 0.80 in 14 comparisons per work, while 16 comparisons on average are required for random CJ (**Figures 3A**, **4A**). In total, with the proposed selection rule 141 fewer comparisons are needed to reach true reliability of 0.80. For essay set 2, with the proposed selection rule on average 3 comparisons per work less are required as compared to random CJ (**Figures 3B**, **4B**). Then, 225 fewer comparisons are needed. Note that the gain in true reliability of the novel progressive selection rule is only moderate with respect to random CJ (0.02-0.03). This can be explained by the relatively large essay sets and the small standard deviations of the true quality levels (**Table 1**; Rangel-Smith and Lynch, 2018; Crompvoets et al., 2020).

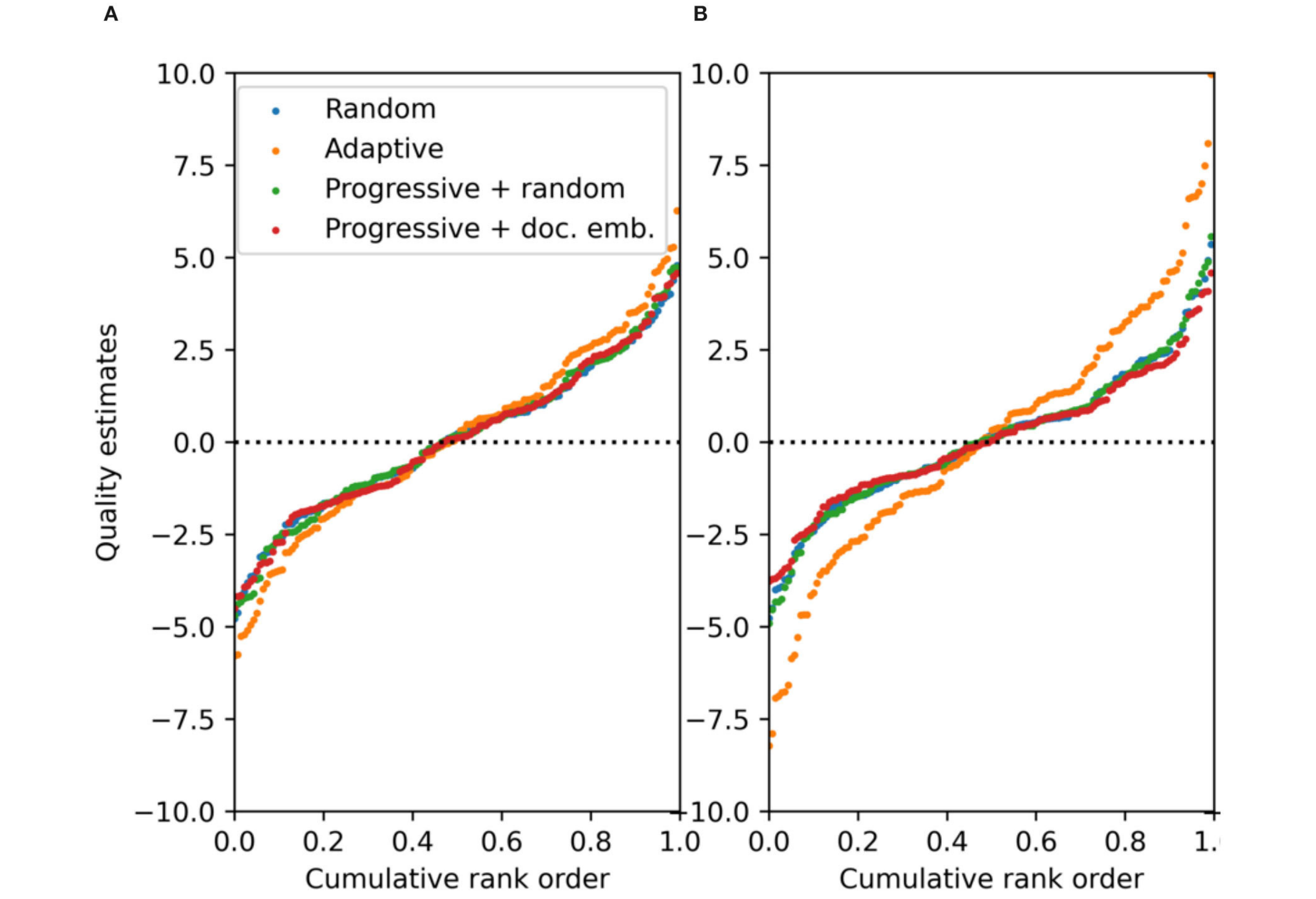


FIGURE 5 | For essay set 1 (N1 = 141), quality parameter estimates with respect to their cumulative ranking for 5 **(A)** and 10 **(B)** Comparisons per work on average. This is assessed for different selection rules: random (CJ), adaptive (ACJ), and a progressive selection with a random component and with a similarity component using document embeddings (100 simulations).

3.3. Evaluation of the Quality Parameter Estimates

To investigate the performance of the SSR estimator we focus on the spread of the quality estimates on the scale and their precision (i.e., uncertainty) (Equation 7). Only the results obtained using the document embeddings as the text representation technique are considered because the SSR results (refer to above) were best for both essay sets. Again random CJ, ACJ, and the progressive selection rule with a random component serve as baselines for comparison.

3.3.1. Spread of the Quality Estimates

Because the absolute differences in quality estimates can vary, the cumulative ranking of the estimates is evaluated, for different average numbers of comparisons.

For five comparisons per work on average, all selection rules result in equivalent estimated quality parameters given their ranking (**Figures 5A**, **6A**). For 10 comparisons per work on average, the differences in estimated quality parameters between ACJ and the other selection rules become noticeable (**Figures 5B**, **6B**). ACJ tends to produce quality estimates that are more spread out than the other selection rules. For ACJ ∼20% of the highest and lowest ranking works will have estimated qualities greater than ±3. For the other selection rules, this is only the case for 5% of the most extreme quality parameter values. The inflated spread of the quality parameter estimates can explain the inflation of the SSR for ACJ (Equation 7). The higher the inflation of the spread of the quality parameter estimates, the more biased the estimates can be. Moreover, when comparing the results of set 1 (**Figure 5**) with the results of set 2 (**Figure 6**), there seems to be an inverse relation between the spread of true quality levels (**Table 1**) and the spread of the estimated quality parameters for ACJ. Namely, the smaller the spread of the true quality levels, the larger the inflation of the spread of the quality parameter estimates, and therefore, the larger the SSR bias for ACJ will be.

3.3.2. The Precision of the Quality Estimates

As the spread of the quality estimates differs between selection rules (**Figures 5**, **6**), the parameter uncertainty is assessed with respect to the cumulative rank order. It can be seen that quality

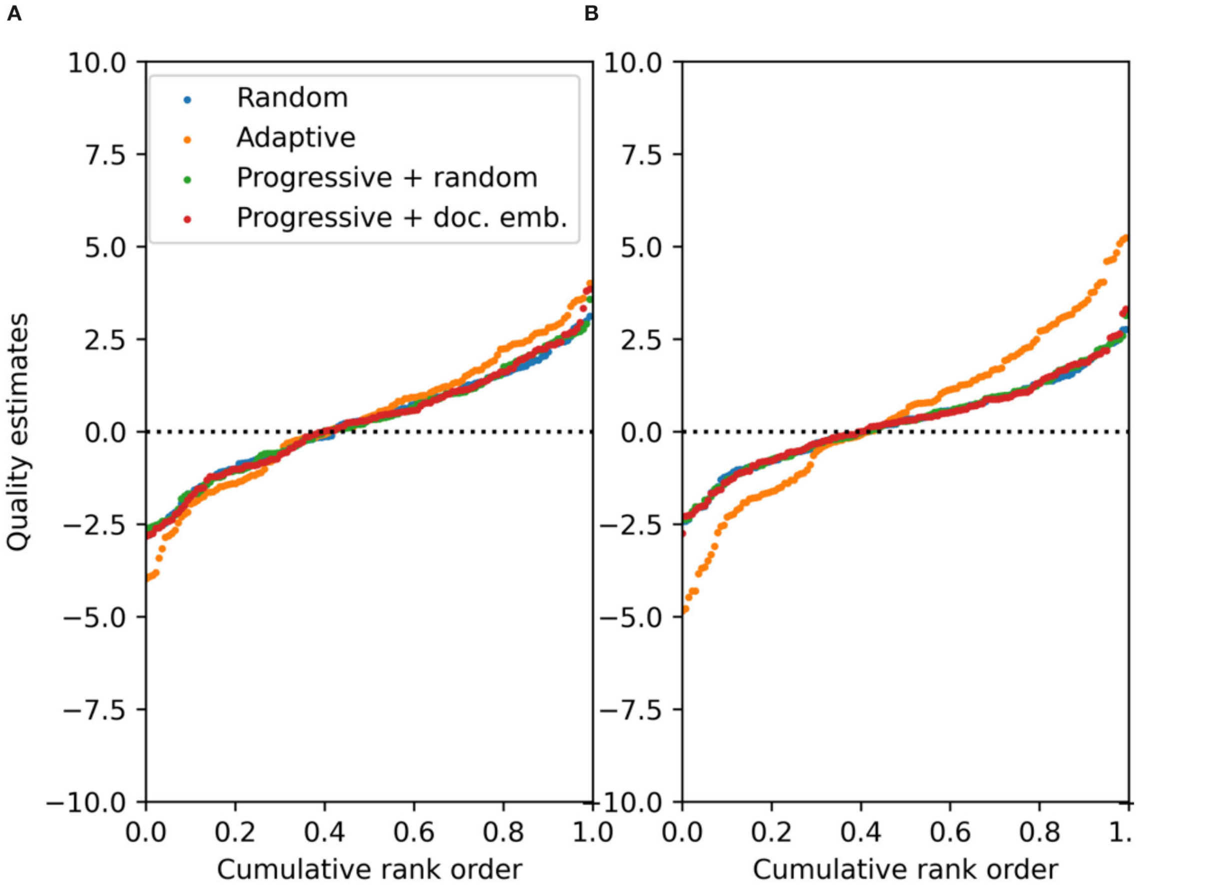


FIGURE 6 | For essay set 2 ($N_2 = 150$), quality parameter estimates with respect to their cumulative ranking for 5 **(A)** and 10 **(B)** comparisons per work on average. This is assessed for different selection rules: random (CJ), adaptive (ACJ), and a progressive selection with a random component and with a similarity component using document embeddings (100 simulations).

parameters are estimated most precisely for middle-ranked essays (**Figures 7**, **8**). This can be explained by the fact that most essay parameters are located around the median. On the other hand, the highest and lowest ranking essay qualities are estimated with less precision. The precision difference between extreme and middle-ranked essays reduces as the average number of comparisons per work increases. This decrease is stronger for essay set 1 (**Figure 7**), which has a larger SD of the true qualities than essay set 2 (**Table 1**). However, for both essay sets ACJ results in more precise quality parameter estimates for 10 or more comparisons per work on average (**Figures 7B**, **8B**). The smaller standard errors for ACJ can inflate the SSR (Equation 7). Note that the increase in precision in ACJ is in itself a desired property; it is its high bias in quality parameter estimates that is undesirable. As the average number of comparisons per work increases, the parameter uncertainty becomes similar for all selection rules (**Figures 7**, **8**). But even then, random CJ results in more uncertain parameter estimates than ACJ.

The progressive selection rule with a similarity component based on document embeddings can show improvements upon random CJ in terms of the precision of the quality parameter estimates. Namely, for essay set 1 a lower uncertainty for high and lower ranked works is obtained after 10 comparisons on average (**Figure 7B**). With respect to the progressive rule with a random component, there is a visible gain in precision for the estimation of quality parameters. For essay set 2, the differences in uncertainty are small (**Figure 8**). This may be explained by the smaller spread of the true quality levels of essay set 2 (**Table 1**).

In sum, the new progressive rule with a similarity component (based on document embeddings), unlike ACJ, does not show inflation of the spread of the quality estimates (**Figures 5**, **6**). This is also observed for the progressive rule with a random component. However, the progressive rule with a similarity component can result in more precise quality parameter estimates than with a random component (**Figures 7**, **8**). This is most notably the case for essay set 1 where the spread of the true quality levels is larger (**Table 1**). For true quality levels that are more spread out a high, unbiased SSR can be obtained with the progressive selection rule based on document embeddings (**Figures 4A**, **3A**) without inflating the spread of the scale of quality parameter estimates (**Figure 5**) and while increasing the precision of the quality parameter estimates (**Figure 7**).

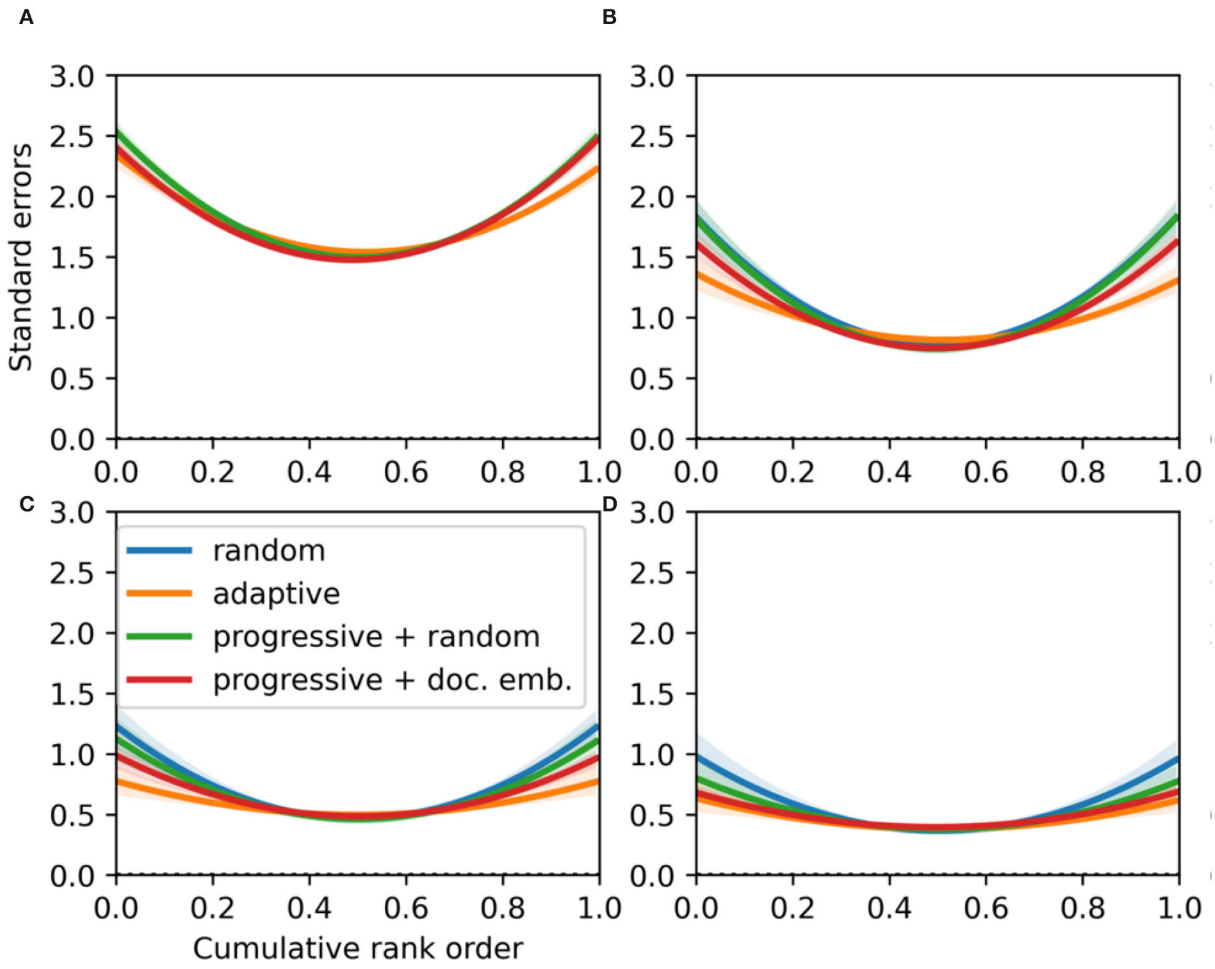


FIGURE 7 | For essay set 1 ($N_1 = 141$), the precision of quality parameter estimates with respect to their cumulative ranking for different averages of work comparisons: 5 **(A)**, 10 **(B)**, 20 **(C)**, and 30 **(D)**, respectively. This is assessed for different selection rules: random (CJ), adaptive (ACJ), and a progressive selection with a random component and with a similarity component using document embeddings (100 simulations). The solid lines indicate the mean values and the transparent bands indicate the 95% point-wise CIs.

4. DISCUSSION

With the proposed selection rule, the essays were initially paired based on the cosine similarities of their vector representations. After the initial phase, the ACJ selection criterion progressively weighted higher in the selection rule (Equation 10). Even though the gain in SSR and true reliability was small, an improvement in terms of SSR estimates and its bias were observed when compared to CJ, ACJ, and a progressive selection rule with a random component. Hence, the proposed selection rule reduced the number of comparisons needed to obtain reliable quality estimates for the essays. The progressive selection rule with a similarity component based on document embeddings performed consistently better than any other selection rule for the two different essay sets. Most importantly, this progressive rule with a similarity component resulted in higher true reliability than the progressive rule with a random component while still reducing the SSR bias quickly. Thus, there is not only evidence that one can alleviate the cold-start by using a progressive selection rule based on the cosine similarities, but also that one can improve the true reliability and the SSR with this selection rule. However, the results indicate the importance of selecting the most appropriate essay representation technique, which was found to be the document embeddings (Le and Mikolov, 2014). The document embeddings were initialized by a pre-trained corpus of word embeddings. A limitation of the simulation design is that in practice multiple raters can compare the same pair while in our design the restriction of one comparison per pair was held. We do not except that by elevating this restriction the results of the proposed progressive selection rule relative to the baseline selection rules would be very different.

Crompvoets et al. (2020) selected essays to be judged that have parameter estimates with the largest standard errors. It was observed that when selecting essays to be judged (work *i*) that way, a large discrepancy occurs in the number of comparisons per essay. Essays with extremer parameter estimates would

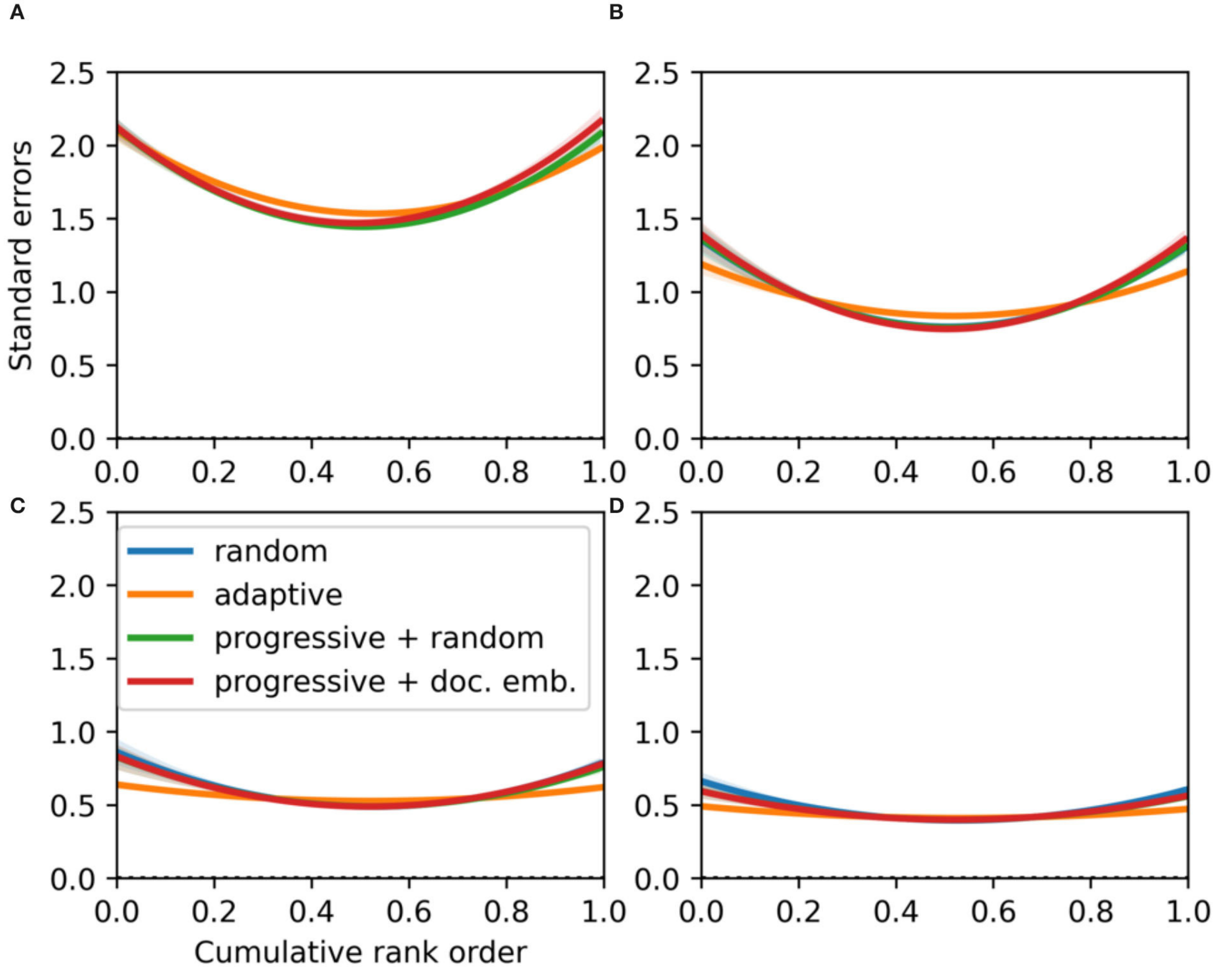


FIGURE 8 | For essay set 2 ($N_2 = 150$), standard error of quality parameter estimates with respect to their cumulative ranking for different averages of work comparisons: 5 **(A)**, 10 **(B)**, 20 **(C)**, and 30 **(D)**, respectively. This is assessed for different selection rules: random (CJ), adaptive (ACJ), and a progressive selection with a random component and with a similarity component using document embeddings (100 simulations). The solid lines indicate the mean values and the transparent bands indicate the 95% point-wise CIs.

consistently be selected as the essay qualities are almost always more uncertain. Instead in this study, it was opted to select the essay to be judged based on the minimal number of times it has been judged. Note that the number of comparisons is also related to the standard errors of the parameter estimates: the standard errors decrease with the number of comparisons (Equation 6). Our approach reflects more practical assessment situations where having an equal amount of comparisons for all works may be preferred. It can be seen as unfair by assessors and students if one essay would be compared more often than another. From a statistical point of view, however, targeting the essays to be judged based on the maximal uncertainty of the parameter estimates may increase the precision of the quality estimates and the SSR even further. Therefore, the selection rule proposed in this study may be improved upon by selecting every essay to be judged (work i) based on the maximal standard error of its parameter estimate. Future research is required with respect to the effects of selecting the essays to be judged based on a combination of the number of times it has been compared and their parameter uncertainty. By doing so, one can prevent too large discrepancies in the number of comparisons per essay while still improving the SSR.

It is expected that for smaller essay sets, the benefits of the progressive selection rule with a similarity component over random CJ will become more apparent. Crompvoets et al. (2020) and Bramley and Vitello (2019) observed that for smaller samples, ACJ can result in a higher gain in the precision of quality parameters and the reliability than random CJ. Furthermore, ACJ can perform well when there is more spread in the true quality levels of works ($\sigma > 2$) (Rangel-Smith and Lynch, 2018). The current results showed that the novel selection rule can produce high true reliability without an increase in SSR bias. Given these results, it is expected that with the proposed selection rule a higher SSR with a small bias can be obtained when it is tested on smaller sample sizes than in the current study. Such cases would represent small classroom assessment situations. Note that document embeddings can be used for smaller essay sets as they can be initiated by a pre-trained corpus of word embeddings (Oostdijk et al., 2013). It is also expected that the benefits would be greater for essay sets that show more high similarities or similarities with more variance. Then more informative initial pairs could be selected. For this study, the essay representations showed rather low similarities (refer to **Table 2**).

As opposed to alleviating the cold-start of ACJ, one can also improve the ACJ-algorithm itself. The proposed progressive selection rules implement the stochastic approach of ACJ from Crompvoets et al. (2020). For an essay to be paired with another, an essay will be selected based on its density value for the distribution of the essay quality estimate that is to be compared (work *i*). That way, the uncertainty of the quality estimate of the essay that is compared is taken into account. However, this assumes that all other essay quality estimates (every work *j*) are deterministic. In order to take the uncertainty of all essay quality estimates into account, a different approach of adaptive pairing is required. A Bayesian adaptive selection rule as proposed in Crompvoets et al. (2021) takes the parameter uncertainty of both work *i* and *j* into account. Every work *i* and *j* are sampled from the conditional posterior distribution of their quality parameter. In the context of item response theory, Barrada et al. (2010) have summarized multiple selection rules that integrate over the weighted likelihood function of an ability parameter: e.g., the Fisher information weighted by the likelihood function or the Kullback-Leibler function weighted by the likelihood function. It is expected that the progressive selection rule with a similarity component would benefit from such a redefined ACJ selection rule.

5. CONCLUSION

The objective of this study was to alleviate the cold-start problem of adaptive comparative judgments, while simultaneously minimizing the bias of the scale separation coefficient that can occur (Bramley, 2015; Rangel-Smith and Lynch, 2018; Bramley and Vitello, 2019; Crompvoets et al., 2020). We proposed the use of text mining as it is possible to extract essay representations before the judgment process has started. A variety of essay representation techniques were considered: term frequency-inverse document frequency, averaged word embeddings, and document embeddings (Aizawa, 2003; Mikolov et al., 2013; Le and Mikolov, 2014). Subsequently, the representations of essays were used to select initial pairs of essays that have high cosine similarities between their representations. Progressively, the selection rule will be more determined by the closeness of the quality estimates given the parameter uncertainty. The simulation results showed that the progressive selection rule can minimize the bias of the scale separation coefficient while still resulting in high true reliability. Out of all representation techniques, the document embeddings of the essays (as initialized by pre-trained word embeddings) consistently showed the best results in terms of scale separation reliability. Moreover, the proposed progressive rule prevents the inflation of the variability of the quality estimates, and it can reduce the uncertainty of the quality estimates—especially for low and high quality essays when the variability of the true quality levels is high. Although the gain in reliability and parameter precision was moderate, it is expected that this gain will be larger for smaller essay sets that show more variability in the true essay qualities and for essays that show more high similarities. A practical example would be its use in classroom assessment contexts.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: the data was previously used for commercial purposes. Requests to access these datasets should be directed to info@comproved.com.

AUTHOR CONTRIBUTIONS

MD and DD: conceptualization and presentation of the problem and design of the simulation study. MD: execution and analysis of the simulation and writing—original draft preparation. MD, DD, and WV: writing—review and editing. DD and WV: supervision. This article originated from the Master thesis MD wrote under the supervision of WV and DD (De Vrindt, 2021). All the authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

The company Comproved was thanked for allowing this study by providing essay texts and scores.

REFERENCES

Ai, Q., Yang, L., Guo, J., and Croft, W. B. (2016). *Analysis of the Paragraph Vector Model for Information Retrieval*. New York, NY. doi: 10.1145/2970398.2970409

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Inform. Process. Manage*. 39, 45–65. doi: 10.1016/S0306-4573(02)00021-3

Barrada, J. R., Olea, J., Ponsoda, V., and Abad, F. J. (2010). A method for the comparison of item selection rules in computerized adaptive testing. *Appl. Psychol. Measure*. 34, 438–452. doi: 10.1177/0146621610370152

Bradley, R. A., and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* 39, 324–345. doi: 10.1093/biomet/39.3-4.324

Bramley, T. (2015). *Investigating the Reliability of Adaptive Comparative Judgment*. Tech. rep., Cambridge Assessment.

Bramley, T., and Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assess. Educ*. 26, 43–58. doi: 10.1080/0969594X.2017.1418734

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Appl. Measure. Educ*. 24, 1–21. doi: 10.1080/08957347.2011.532417

Coenen, T., Coertjens, L., Vlerick, P., Lesterhuis, M., Mortier, A. V., Donche, V., et al. (2018). An information system design theory for the comparative judgement of competences. *Eur. J. Inform. Syst*. 27, 248–261. doi: 10.1080/0960085X.2018.1445461

Crompvoets, E. A., Béguin, A. A., and Sijtsma, K. (2020). Adaptive pairwise comparison for educational measurement. *J. Educ. Behav. Stat*. 45, 316–338. doi: 10.3102/1076998619890589

Crompvoets, E. A. V., Beguin, A., and Sijtsma, K. (2021). *Pairwise Comparison Using a Bayesian Selection Algorithm: Efficient Holistic Measurement*. doi: 10.31234/osf.io/32nhp

Davey, T., Nering, M. L., and Thompson, T. (1997). *Realistic Simulation of Item Response Data, Vol. 97*. Iowa City, IA: ERIC.

De Vrindt, M. (2021). *Text mining to alleviate the cold-start problem* (Master's thesis). KU Leuven, Leuven, Belgium.

Hunter, D. R. (2004). MM Algorithms for generalized Bradley Terry models. *Ann. Stat*. 32, 384–406. doi: 10.1214/aos/1079120141

Jones, I., Bisson, M., Gilmore, C., and Inglis, M. (2019). Measuring conceptual understanding in randomised controlled trials: can comparative judgement help? *Br. Educ. Res. J*. 45, 662–680. doi: 10.1002/berj.3519

Jones, I., and Inglis, M. (2015). The problem of assessing problem solving: can comparative judgement help? *Educ. Stud. Math*. 89, 337–355. doi: 10.1007/s10649-015-9607-1

Lau, J. H., and Baldwin, T. (2016). "An empirical evaluation of doc2vec with practical insights into document embedding generation," in *Proceedings of the 1st Workshop on Representation Learning for NLP* (Berlin: Association for Computational Linguistics), 78–86. doi: 10.18653/v1/W16-1609

Le, Q. V., and Mikolov, T. (2014). Distributed representations of sentences and documents. *arXiv [Preprint]*. arXiv: 1405.4053. doi: 10.48550/arXiv.1405.4053

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

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv [Preprint]*. arXiv: 1301.3781. doi: 10.48550/arXiv.1301.3781

Oostdijk, N., Reynaert, M., Hoste, V., and Schuurman, I. (2013). "The construction of a 500-million-word reference corpus of contemporary written Dutch," in *Essential Speech and Language Technology for Dutch*, eds P. Spijns and J. Odijk (Berlin; Heidelberg: Springer), 219–247. doi: 10.1007/978-3-642-30910-6_13

Pollitt, A. (2004). "Let's stop marking exams," in *IAEA Conference* (Philadelphia, PA).

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assess. Educ*. 19, 281–300. doi: 10.1080/0969594X.2012.665354

Rangel-Smith, C., and Lynch, D. (2018). "Addressing the issue of bias in the measurement of reliability in the method of adaptive comparative judgment," in *36th Pupils' Attitudes towards Technology Conference* (Athlone), 378–387.

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

Singh, R., and Singh, S. (2021). Text similarity measures in news articles by vector space model using NLP. *J. Instit. Eng. Ser. B* 102, 329–338. doi: 10.1007/s40031-020-00501-5

Thurstone, L. L. (1927). The method of paired comparisons for social values. *J. Abnorm. Soc. Psychol*. 21:384. doi: 10.1037/h0065439

Tulkens, S., Emmery, C., and Daelemans, W. (2016). Evaluating unsupervised Dutch Word embeddings as a linguistic resource. *arXiv [Preprint]*. arXiv: 1607.00225. doi: 10.48550/arXiv.1607.00225

Verhavert, S., De Maeyer, S., Donche, V., and Coertjens, L. (2018). Scale separation reliability: what does it mean in the context of comparative judgment? *Appl. Psychol. Meas*. 42, 428–445. doi: 10.1177/0146621617748321

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 08 September 2022
DOI 10.3389/feduc.2022.826742

Pairwise comparison scale extension using core linking sets

Stephen Humphry*† and Ken Bredemeyer†

Graduate School of Education, University of Western Australia, Perth, WA, Australia

Check for updates

OPEN ACCESS

EDITED BY
Tine Van Daal,
University of Antwerp, Belgium

REVIEWED BY
Dries Debeer,
Ghent University, Belgium
Soo Lee,
American Institutes for Research,
United States

*CORRESPONDENCE
Stephen Humphry
stephen.humphry@uwa.edu.au

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement,
a section of the journal
Frontiers in Education

RECEIVED 01 December 2021
ACCEPTED 17 August 2022
PUBLISHED 08 September 2022

CITATION
Humphry S and Bredemeyer K (2022) Pairwise comparison scale extension using core linking sets.
Front. Educ. 7:826742.
doi: 10.3389/feduc.2022.826742

Pairwise comparisons can be used to equate two sets of educational performances. In this article, a simple method for the joint scaling of *two or more* sets of assessment performances is described and illustrated. This method is applicable where a scale of student abilities has already been formed, and the scale is to be extended to include additional performances. It requires a subset of already scaled performances, which is designated as a core linking set for the purpose of scale extension. The article illustrates the application of the method to construct a scale with a larger range of latent abilities, using fewer additional comparisons compared to the standard method of pairwise comparisons. The design differs from standard pairwise comparisons in the way performances are paired. The method of pairing performances can also be used to efficiently place individual performances on an existing scale.

KEYWORDS

pairwise comparison, comparative judgment, linking set method, equating, efficiency

Introduction

Pairwise comparisons can be used to assess students' work, such as essays and language tests (Heldsinger and Humphry, 2010; Humphry and McGrane, 2015; Steedle and Ferrara, 2016; Humphry and Heldsinger, 2019, 2020), as a replacement for rubric marking (Pollit, 2009, 2012; Steedle and Ferrara, 2016). The method of pairwise comparisons can also be used to equate two sets of performances without requiring common items or common persons (using instead common judges). For example, it can be used to equate a scale obtained from one rubric to the scale obtained from another, through comparisons of performances on the two scales (Humphry and McGrane, 2015). This type of equating design cannot be achieved without the use of pairwise comparisons or a similar method.

Using pairwise comparisons for equating two sets of assessment performances is well-documented in the literature. This article introduces and illustrates a simple method for scale extension in contexts where one set of performances has already been scaled and another set of performances is equated with the scaled set through joint scaling. The method enables researchers to concentrate pairings to efficiently align scales formed from separate sets of performances and it also affords other advantages.

Because it connects the two data sets in the equating design to enable joint scaling, the method introduced in this article requires the selection of a set of already scaled performances, which is referred to as the *core* set. Then performances on the new scale, which are to be equated onto an existing scale, are compared against the core set. In the generation of pairs, these performances are referred to as *non-core* and it follows that all the comparisons used to connect the scales are core vs. *non-core*. The relevance of *core* and *non-core* sets is most clearly apparent when there are at least three sets where two or more *non-core* sets are placed on a common scale through a core set.

The aim of the article is not to study application of the method under a range of conditions; rather the scope is limited to a single empirical application and a single simulation study. The introductory context is chosen to highlight general considerations for application of the method.

In addition to scale extension, pairwise comparisons using pairs generated as core vs. non-core can also be applied *post-hoc* to efficiently place new performances on an existing scale when those new performances have not been scaled. To place performances on an existing scale, the core set would be drawn from already scaled performances and the performances to be placed on the scale designated non-core. This application is discussed later in the article, but is not its main focus. Nevertheless, we discuss implications for future research, including the application of computer adaptive presentation of pairs based on existing calibrated performance banks.

The structure of this article is as follows. First, a brief background to the method of pairwise comparisons and its relevance in educational assessment is presented. Next, a design and method for equating two separate scales using core vs. non-core pairs is detailed. The method is demonstrated using empirical data collected from a persuasive writing task, and then applied in a simulation study. The aim of the empirical study is to extend a writing scale formed on the basis of paired comparisons, and subsequently to obtain performance exemplars for use by teachers in separate assessments of their own students' performances. The aim of the simulation study is to emulate the empirical study, to ascertain the effectiveness of the method used to extend the scale, where data fit the relevant model. In the empirical study, the writing task was administered to primary school and secondary school students, whose performances were judged, using pairwise comparisons, by experienced markers using an online platform. The estimation procedure for placing the performances on a scale of latent writing ability is outlined for both the empirical data from the school assessment task and the simulation study. The resulting scales are evaluated using fit statistics and, for the simulation study, by comparing the estimated and simulated parameters. Lastly, a discussion follows which includes the benefits of the method, considerations for its application, and limitations of the studies presented.

Background

As broader background, the method of pairwise comparisons is based on Thurstone's law of comparative judgment (Thurstone, 1927). Bradley and Terry (1952), and later Luce (1959), showed that Thurstone's equations for the analysis of pairwise comparison data could be simplified using the cumulative logistic function. The resulting Bradley-Terry-Luce (BTL) model is used to estimate the latent ability of the persons in this study. The BTL model has the same form as the Rasch model (Andrich, 1978), but the probabilities of success are defined using the differences between performance estimates, rather than using the differences between ability estimates and item difficulties.

The BTL model defines the probability that performance *a* is compared favorably over performance *b* as follows:

$$P\left(a > b\right) = \frac{e^{a-b}}{1 + e^{a-b}}$$

where *a* and *b* are the parameters denoting the latent writing abilities inferred from the quality of performances. As with Rasch modeling, the BTL model provides a scale for performances (provided there are enough comparisons) if the data fit the model adequately.

An excellent and more detailed discussion of the background into the method of pairwise comparisons can be found in Bramley (2007). Bramley's article covers the development of pairwise comparison methodology from the adaptation of Thurstone's original work to the form used in the current study. See also Humphry and Heldsinger (2019) for a brief overview of some key literature focusing on different aspects of the application of pairwise comparisons in education.

Pairwise comparisons offer a very flexible design for parameter estimation. It is not necessary to compare each performance with every other performance. Pollit (2012, pp. 160) states that this "system is extraordinarily robust." This means that sparse data can be analyzed to yield performance locations with acceptable standard errors of estimation.

To obtain sufficiently accurate locations using pairwise comparisons, it is useful to specify the number of times each performance is compared to others. If a performance is compared too few times, the standard error of estimation will be high, so there will be a large degree of uncertainty in the location of the performance. Various authors have offered recommendations for the minimum number of comparisons generally required (Verhavert et al., 2019). Pollit (2012, pp. 160) claims that, "if every object is compared about 10 times to suitable other objects, this will generate a data set that is adequate to estimate the values of every object on a single scale." This is also a key consideration for joint scaling of sets of performances on existing scales, as elaborated later.

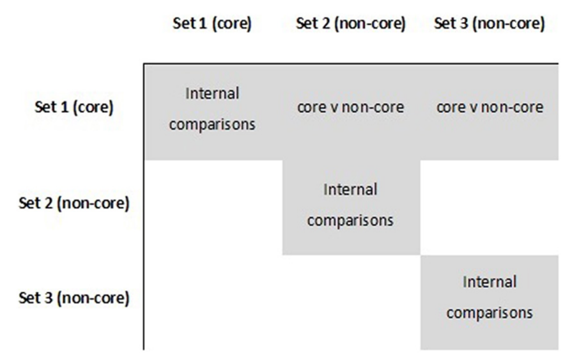
Pollit (2012) refers to the concept of *chaining* performances to reduce the time spent judging comparisons. In the moderation exercise presented in Pollit (2012), each pair of two successive comparisons contains a common performance, so that reading time is reduced on the second comparison. In the current study, common performances are included for more than two consecutive comparisons to further improve judging efficiency. The number of consecutive comparisons which contain a common performance is referred to in this article as the "chaining constant." Although Pollit's main reason for including chaining of performances in the design is to improve time efficiency, it stands to reason that the cognitive load for judges is also reduced because they do not need to become familiar with two new performances each and every time they see a new pair. Chaining performances in consecutive comparisons has some potential to introduce violations of the statistical assumption of independent comparisons, though Pollit (2012) notes that no evidence of chaining bias has yet been found.

Due to the robustness and flexibility of the pairwise comparison method, judgments of pairs generated using the core vs. non-core method can be combined with judgments of standard pairs and the BTL model applied, provided there is a core linking set and the comparisons were made using the same judging criteria. By combining core vs. non-core comparisons with standard comparisons, a new scale can be obtained for the new performances that is anchored to the existing scale.

The use of core vs. non-core comparisons is alluded to by Steedle and Ferrara (2016, p. 211) in stating: "if desired, these [pairwise] estimates can be anchored to a rubric scale by including anchor papers with fixed scores in the judgment and estimation process." The procedure described by Steedle and Ferrara is equivalent to a design that includes standard pairs plus core vs. non-core pairs, as described in this article.

In this article, OUTFIT MNSQ is used as an indicator of model fit to check the fit of the performances. The OUTFIT MNSQ statistic is computed in the same way as in applications of the Rasch model (Wright and Stone, 1979; Wright and Masters, 1982) except that the observed and expected scores are related to two person parameters in the BTL model rather than person and item parameters in Rasch's model. The expected value of the Outfit statistic, or unweighted mean-squared standardized residual, is approximately 1. An often-used range of acceptable limits for the Outfit index is 0.7–1.3 (Smith et al., 2008).

The Person Separation Index is used as an indicator of the internal consistency of the judgments on which the scale is based and is modeled on Cronbach's alpha. Its minimum value is effectively 0 and its maximum is 1. A higher value indicates higher internal consistency. Relevant to the interpretation of results, for a given level of internal consistency, the separation index will be higher if there are more comparisons because there is more Fisher information and smaller standard errors, as touched upon by Heldsinger and Humphry (2013).



	Set 1 (core)	Set 2 (non-core)	Set 3 (non-core)
Set 1 (core)	Internal comparisons	core v non-core	core v non-core
Set 2 (non-core)		Internal comparisons	
Set 3 (non-core)			Internal comparisons

FIGURE 1
Schematic depiction of a data matrix with a core set and two non-core sets.

Materials and methods

Rationale for using core and non-core sets

To explain the core and non-core distinction and the use of core sets for joint scaling in general terms, it is instructive to consider situations in which a core set of performances is used to join three or more separate data sets. **Figure 1** depicts a case of three sets in which the core set links the other two data sets for which there are no direct comparisons between performances. In this case, performances in Sets 2 and 3 will be placed on a common scale only through comparisons with performances in the core set (Set 1) and only if there is sufficient overlap between Sets 1 and 2, and Sets 1 and 3.

Figure 1 depicts the basis of the method using a simple case in which all performances in Set 1 form a core set, all performances in Set 2 form a non-core set, and all performances in Set 3 form a non-core set. To avoid confusion, we note that in the empirical and simulation studies used in this article, the core and non-core sets are subsets of primary and secondary performances, i.e., they are subsets of larger sets. The reasons for selecting subsets for core vs. non-core comparisons are explained to follow.

More generally, the data matrix may comprise any number of sets that have internal comparisons, and in principle the core set will provide a basis for joint scaling on a common scale. Thus, a single core set may be used to equate three, four or more other sets that each have only internal comparisons prior to the use of comparisons with performances in a core set.

The most extreme case is that in which each non-core performance comprises its own set containing just one performance. In this case, comparisons against the core set are the means of placing individual performances on a common scale.

Although the logic of the core and non-core distinction is most apparent when there are at least three sets, there are advantages to pairing performances using the distinction when there are two sets. Further, considerations applicable to two sets are also applicable to cases in which there are three or more sets to be jointly scaled. The empirical and simulation studies described below illustrate the use of core vs. non-core pairings between two sets to obtain the advantages of targeted selection of performances and the availability of specific diagnostic information to evaluate joint scaling. These advantages are discussed in further detail later in the article.

Requirements of joint scaling

To jointly scale all performances by selecting a core set and one or more non-core sets, there needs to be sufficient information from the performances. When the core vs. non-core method is used, there need to be internal comparisons of performances within both the core and non-core sets before the scale locations of performances in the two sets can be equated with each other.

Given limited resources, it may be necessary to concentrate available comparisons on the most useful pairings for linking. To explain the nature of information required for joint scaling, consider an extreme case in which there is no information and joint scaling is not possible. Specifically, let Set 1 be the core set and suppose there is only one performance *j* (non-core) from Set 2 used to equate the two sets, termed the link Subset *L*, and only one comparison of performance *j* in Subset *L* against a performance *i* in Set 1. In this case, performance *j* has an estimate on the scale comprising Set 2 performances but it is not possible to obtain an estimate for performance *j* based on comparisons with Set 1 performances. Therefore, it cannot provide any information to align the two scales.

The first case can be expanded to a case in which there is a significant number of performances in a link Subset *L*, contained within Set 2 used to equate the two sets, but where only one comparison is made between each performance in Subset *L* against a performance in the core Set 1. Using the reasoning above, it is not possible to obtain an estimate for any performance in Set 2 based on comparisons with Set 1 performances. Thus, comparisons for these performances cannot provide any information to align the two scales.

If we further expand the case so that there are at least two comparisons between performances from Set 2 and performances in Set 1, then estimates of Set 2 performances can be obtained on the scale for the Set 1 performances. In this case, comparisons for the performances do provide information to align the scales. However, if Set 2 performances that are compared with Set 1 performances have very few comparisons with Set 1 performances, the standard errors are large. Accordingly, if there is little information and the standard errors are large, plots such as those in Figures 2, 3 are likely to provide little information about whether there is a linear relation between the two sets of location estimates for the Set 2 performances, based on comparisons with Set 1 vs. comparisons with Set 2. With little information, the measurement error will obscure the association. On the other hand, if there is sufficient information, such plots can be expected to provide information about whether there is a linear relationship.

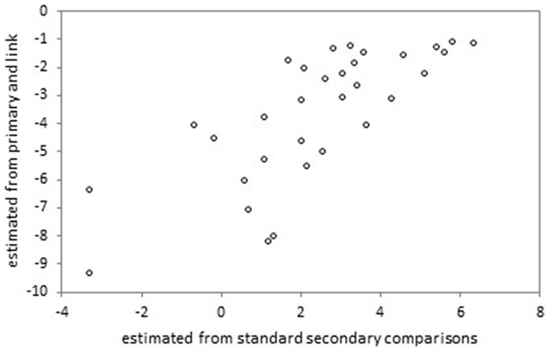


FIGURE 2
Location estimates of the 32 non-core secondary performances, from two independent analyses in the scale extension study.

Design considerations

Following from the considerations detailed above, the optimal design of a scale extension paired comparison exercise depends on factors that include: (i) the number of performances in sets; (ii) the number of new comparisons that can be made with available resources; and (iii) the abilities of students producing different sets of performances.

If it is possible to make enough comparisons such that random pairings ensure performances in Set 2 are compared a reasonable number of times against performances in Set 1 (say, more than seven times) then this option can be used and diagnostic information will be useful. Given the numbers of performances in Sets 1 and 2 and the available resources for comparisons, if the number of comparisons of a Set 2 performance against a Set 1 performance is typically low with random pairings, then the core vs. non-core pairing method provides advantages. The advantages, relevant to the empirical illustration of the method, are detailed later in this article.

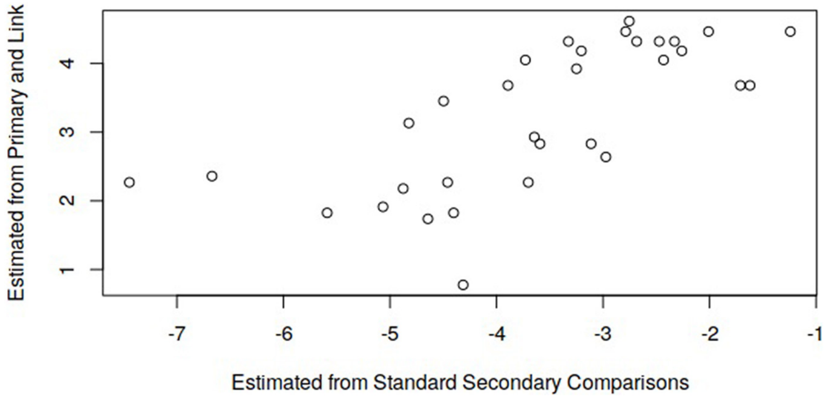


FIGURE 3
Location estimates of the link set from scaling of independent data sets in the simulation study.

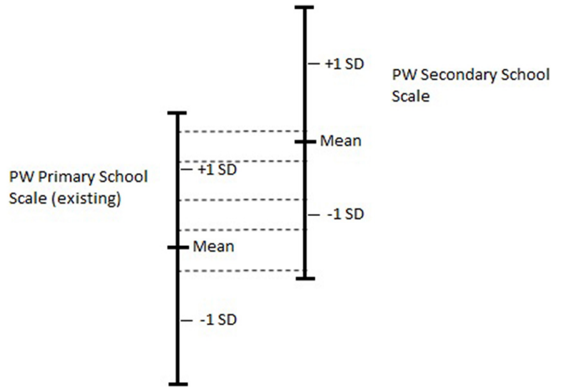


FIGURE 4
Schematic showing the scale extension design. Vertical lines represent the extent of the ability scales.

Paired comparison design for illustrative study

Figure 4 depicts the design of the empirical and simulated studies. In the figure, the horizontal dashed lines represent the pairings of performances to connect the two scales. They also convey a range in which the core vs. non-core comparisons are generated. The primary scale obtained from pairwise comparisons (PW Primary School Scale) is the existing scale, depicted on the left-hand-side. The secondary scale obtained from pairwise comparisons (PW Secondary School Scale) is depicted on the right-hand-side.

Empirical study

The empirical study described in this article focuses on the extension of a primary school Writing scale, whose construction is described in Humphry and Heldsinger (2019). For the scale extension project, a scale of latent writing ability was already formed using pairwise comparisons of primary school performances. The goal of the empirical study was to extend the pre-existing scale upwards to include performances of secondary school students in years 7–9. It was assumed, prior to equating, that the secondary school performances would be generally of a higher quality than the primary school performances, but there would be sufficient overlap in the quality of performances from the two groups to enable this type of equating.

Primary school standard comparisons

Located on the existing primary school scale were 162 writing performances from primary school students. The construction of the scale, including the data collection, test administration, judgments, and pairwise comparison procedure are detailed in Humphry and Heldsinger (2019, see pp. 509–510). The criteria for making judgments as to which performance is better in each pair is also detailed in Humphry and Heldsinger (2019). In the study, a persuasive writing task was administered by classroom teachers, who had been provided with instructions and a choice of topics to present to their students. A total of 3,228 pairs were compared by 18 judges who were all experienced classroom teachers.

Secondary school standard comparisons

To scale the secondary performances, 111 secondary school performances from students in years 7–9 were compared with one another by 16 judges. The judges made a total of 1,018 comparisons, with most judges making 60 comparisons each. Of the 16 judges, five were highly experienced assessors of both primary and secondary Writing, one was a primary classroom teacher, and the rest were secondary classroom teachers.

Core vs. non-core comparisons

The current empirical study, designed to place the two sets of performances on the same scale, involved pairing primary school performances to secondary school performances to form the core vs. non-core pairs. Of the 162 primary and 111 secondary school performances, 82 primary and 32 secondary school performances were selected to be used in the core vs. non-core comparisons. To obtain performances with an overlapping range of performance levels, primary school performances with the highest locations and secondary performances with the lowest locations were selected.

A total of 2,624 core vs. non-core pairs were generated and allocated to judges. Four judges, who were very experienced in marking both primary and secondary Writing, made 656 comparisons each, resulting in all the core vs. non-core pairs generated being compared.

Generation of core vs. non-core comparisons

For the purpose of core vs. non-core comparisons, pairs were generated between the two sets of performances and not within either set. Selected primary school performances were designated as core and selected secondary school performances were designated as non-core. The top 82 primary school performances and the bottom 32 secondary school performances were selected into these sets, based on estimated locations from standard pairwise scaling of the primary and secondary performances separately. The method generally aims to place non-core performances on the scale formed using the core, linking set of performances. The key requirement for pair generation using this method is to specify the number of times each non-core performance is included in the set of pairs allocated to judges. In the empirical project, each non-core (secondary school) performance was included exactly 82 times. For each pairing, a core performance is matched randomly with a non-core performance (without replication). Random sampling without replacement was used in the pairing procedure, given other applicable constraints on pair generation, in order to ensure that performances were sufficiently connected for joint scaling. The pairs were generated using the pair generator R package (Bredemeyer, 2021a).

Pair presentation

Pairs of performances were presented side by side to judges to make comparisons using online software. The left vs. right presentation on the screen was fully randomized for the performances. Each performance was included in a comparison 22 times on average, and a chaining constant of four was used to reduce the cognitive load of judges.

Scaling and scale extension for the empirical component

To jointly scale the primary and secondary scales in both simulation and empirical studies, comparisons from the three sets of judgments—primary school standard comparisons, secondary school standard comparisons, and core vs. non-core comparisons—were combined. The combined set of comparisons were used to estimate the abilities of performances based on the Bradley-Terry-Luce model, which is implemented in the PairwiseComparisons R package (Bredemeyer, 2021b) built in the R statistical and programming environment (R Core Team, 2021). Scale locations for each performance were obtained using an estimation algorithm that calculates the performance location in logits, centered on zero. For the applied study, a shift constant was added to all performance locations, so that locations were centered on the primary school performances (so that the mean of the primary school locations was zero). Applying the shift constant simply aligns the combined scale to the original scale of the primary school performances.

Scaling the three sets of pairwise comparisons together ensures that the origin of the scale is consistent for all performances. In summary, the steps for the joint scaling of performances were as follows. First, the primary school performances were scaled using standard pairs, in which all pairs were sampled from a list of all possible pairs of primary school performances. Second, secondary school performances were scaled also based on standard pairs. Third, a set of the primary performances with the highest scale locations was selected and a set of the lowest secondary performances was selected. Fourth, primary school and secondary school performances were compared using core vs. non-core pair generation and presentation of the pairs to judges for comparison. Once the pairwise comparisons had been made, all data were combined in a single data set and abilities were estimated using the BTL model. As a last step in the empirical study, to align the final scale with the original primary scale, a shift was applied such that the primary school performances have the same location as they did in the original primary scale.

Simulation study

Simulation specifications and details

Simulation specifications were chosen to emulate the Writing scale equating project in terms of the numbers of performances and the direction of the difference between the means. The specifications in **Table 1** were followed for the simulation study so it matches the empirical study. Because the goal of the empirical study was to align two previously formed scales, the set of comparisons of primary school performances and the set of comparisons of secondary school performances were held constant over multiple repetitions of the simulation to emulate the design of the empirical study. The core vs. non-core comparisons were generated uniquely over 30 repetitions of the simulation. The top 82 primary school performances and the bottom 32 secondary school performances were selected into the core and non-core sets based on their simulated locations. For each of the 30 repetitions of the simulation, joint scaling of all performances was performed, and a shift constant was calculated in order to center the primary school performance locations on zero.

The number of performances, and the mean and standard deviation of the person locations, specified for the simulation, are shown in **Table 1**. The normal random distribution was used to generate logit locations for both the primary school and secondary school simulated performances, based on the specifications in **Table 1**.

To demonstrate the efficiency of the use of pairings, a similar simulation was conducted in which pairings between performances in the secondary and primary sets were made at random (without replacement). For this simulation, all primary school performances and all secondary performances were in the sampling pools for selection into the core and non-core sets, respectively. This random design simulation was the same as the core vs. non-core simulation in other respects.

Pair generation for the simulated component

For the core vs. non-core simulation, standard pairs were generated for both primary school and secondary school performances using the pair generator R package (Bredemeyer, 2021a). A total of 1,622 pairs were generated for primary school performances and 1,112 pairs were generated for secondary school performances. Each performance was included 40 times on average for the primary school set and 20 times on average in the secondary school set. Primary school and secondary school pairs were formed only once as the standard scales were considered to exist prior to the application of the core vs. non-core method.

Because the core vs. non-core pairings are exhaustive in the empirical data, the core vs. non-core pairs were also formed only once for the simulation; all core performances are compared against all non-core performances, and therefore the comparisons did not vary over repetitions of the simulation. Each non-core performance was paired against every core performance, so that 2,624 comparisons of primary school performances against secondary school performances were made, in each repetition of the simulation.

When all three sets of judgments—primary school standard comparisons, secondary school standard comparisons, and core vs. non-core comparisons—were combined, there was a total of 5,358 comparisons.

Simulated comparisons

Judgments, as to which of the pairs was deemed better, were simulated using the PairwiseComparisons R package (Bredemeyer, 2021b). PairwiseComparisons simulates judgments of pairwise comparisons by generating deviates of the binomial distribution, where the probability of favorably comparing one performance is the probability defined by the BTL model.

Scaling and scale extension for the simulated component

The secondary Writing performances were scaled using the BTL model in the same manner that the primary Writing performances were scaled. To ascertain how well the scales were connected, the mean difference between estimated locations of secondary and primary performances was compared with the difference between the simulated locations of the secondary and primary performances.

The reason for comparing the mean differences is as follows. The simulated difference between the mean secondary and primary locations is 6.75. The estimate of each individual scale location contains measurement error; however, measurement error only has a minor impact on the *mean* scale locations for the primary and secondary person groups. Therefore, if the two scales are aligned, the estimated mean difference between the person groups will be accurate and consistent with the simulated mean difference of 6.75. Thus, the accuracy of the estimation of the mean difference indicates the accuracy of the alignment of the primary and secondary scales.

In addition, if the scales are aligned, the simulated primary and secondary locations will be correlated with the estimated locations. Also, the plot of simulated vs. estimated locations will follow a single line without being disjointed across year groups. A scatterplot showing the correspondence between simulated and estimated locations for both primary and secondary performances is provided in the results to follow (**Figure 5**).

Results

Empirical study

The Person Separation Index of the joint scale was 0.977, indicating a generally high level of internal consistency among the judgments. On the same scale, 58 of the 273 performances had OUTFIT MNSQ values greater than 1.3, indicating that there was a number of performances with relatively poor fit to the model. On the other hand, a relatively large number of performances had OUTFIT MNSQ values below 0.7 ($n = 143$).

Of the link set of performances used to connect the scales, 8 of 82 primary and 3 of 32 secondary performances had OUTFIT MNSQ values greater than 1.3, indicating that the sets were connected by performances that mostly had acceptable fit to the BTL model.

To evaluate whether the scales were connected by performances whose locations had a linear association, the performances of the secondary non-core set were independently scaled based on: (i) primary standard pairs combined with the core vs. non-core comparisons; and (ii) secondary standard

TABLE 1 Mean and standard deviation of the simulated and estimated parameters for the primary and secondary performances in the simulation study.

		Specifications		Estimated (mean)		Estimated (range)	
	N	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
Primary	162	−2.75	4.10	−2.96	4.04	−3.02, −2.90	3.95, 4.14
Secondary	111	4.01	3.26	4.32	3.62	4.24, 4.41	3.59, 3.65

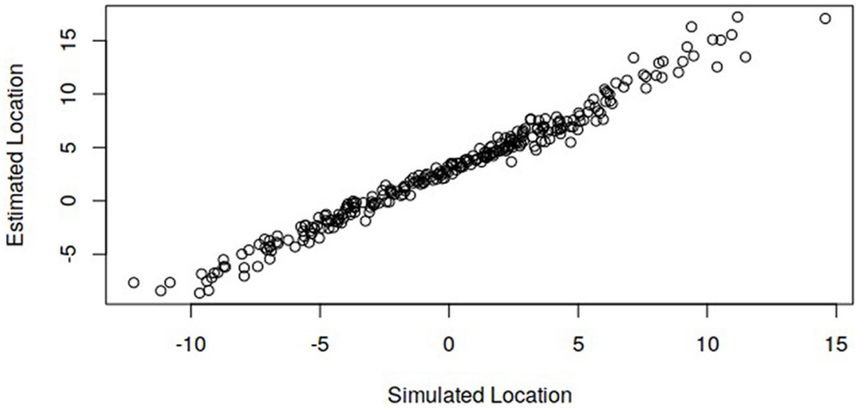


FIGURE 5
Simulated locations against estimated locations based on joint scaling of all 273 performances in the simulation study.

TABLE 2 Mean and standard deviation of location estimates for the primary and secondary performances in the empirical study.

	N	Mean	Std. dev.
Primary	162	−2.75	4.10
Secondary	111	4.01	3.26

pairs. The resulting scatterplot of the two sets of estimates for the 32 secondary non-core performances is shown in Figure 2. The association is reasonably linear with a Pearson correlation of $r = 0.751$.

Table 2 shows the mean and standard deviation of the primary and secondary locations in the empirical study.

Simulation study

The Person Separation Index for the joint scale of the core vs. non-core simulation study was 0.98 for all repetitions of the simulation. The Person Separation Index varied among simulation repetitions only by the third decimal place. This indicates a high level of internal consistency among the judgments. From joint scaling, on average across the 30 repetitions, 27 of the 273 performances had OUTFIT MNSQ values greater than 1.3 and 166 of the 273 performances had OUTFIT MNSQ values less than 0.7. The proportion of OUTFIT MNSQ values above 1.3 is somewhat higher than expected in theoretical terms. However, because the data was simulated strictly according to the model, it is likely that the high proportion is related to the specifics of the design. The information nevertheless provides a reference point for the results in the empirical study with its similar design.

Figure 5 shows the association between the simulated locations and the estimated locations, with the latter based on joint scaling of primary and secondary performances. The locations shown in Figure 5 are for a single repetition of the simulation and are indicative of results obtained in the simulations. It can be seen that the bivariate locations follow a single line, indicating the scales have been aligned such that primary and secondary locations are on the same scale.

Figure 3 shows the estimates of the secondary link performances from independent scaling of the secondary data on the *x*-axis and the primary linking set data on the *y*-axis. The Pearson correlation is $r = 0.698$, indicating a linear association that provides a good basis for connecting the two scales. The locations shown in Figure 3 are for a single repetition of the simulation.

The results from: (i) core vs. non-core pairings; and (ii) random pairings, were compared. The cross-set pairings provide information about secondary estimates relative to the primary estimates only if they have non-extreme scores on the primary scale. In the random pairings design, of the cross-set comparisons and averaged across simulations, 32.5% of pairs were involved in comparisons of secondary performances that had extreme estimates based on comparisons with primary performances. As explained in the justification for the approach, these pairs provide no information about the relation between secondary and primary estimates and are wasted for the purpose of aligning the scales. In the core vs. non-core design, of the cross-set comparisons, none of the pairs were involved in comparisons of secondary performances that had extreme estimates based on comparisons with primary performances; that is, none of the pairs were wasted.

The mean difference of the estimates indicates how well the origins of the scales are aligned with each other based on the comparisons. The mean difference between simulated secondary and primary performances is 6.970 on the common scale. The mean difference is more accurately estimated in the core vs. non-core design (7.273) than the random linking design (7.435). The standard deviation of the mean difference provides an estimate of the uncertainty of the estimate around the mean difference based on multiple simulations. The standard deviation is 2.25 times greater in the random linking design (0.162) than the core vs. non-core design (0.072). This effectively indicates a larger standard error of equating resulting from less information for aligning secondary performances on the primary scale. The estimates of the mean differences, in both designs, are larger than simulated due to some bias in the estimates of performances with the highest and lowest locations on the scale (see **Figure 5**).

The Person Separation Index for the joint scale of the random pairings simulation was 0.98, varying only by the third decimal place among the 30 repetitions of the simulation. From joint scaling, on average across the 30 repetitions, 21 of the 273 performances had OUTFIT MNSQ values greater than 1.3, and 213 of the 273 performances had OUTFIT MNSQ values less than 0.7.

Discussion

The empirical and simulation studies enable discussion of specific considerations applicable to the selection of core and non-core sets for the purpose of scale extension. The considerations apply to cases in which there are two or more non-core sets (which may be subsets of larger sets) that have been scaled based on internal comparisons.

Scaling performances with *core* vs. *non-core* comparisons adds flexibility in relevant contexts because this method takes advantage of a measurement scale already formed using standard pairwise comparisons. As the *core* set of performances have already been scaled, the number of *all* comparisons can be reduced relative to the standard pairwise method.

Using core and non-core sets enables practitioners to more effectively concentrate the use of available pairwise comparisons to achieve joint scaling given finite resources for comparisons. When resources are limited beyond a certain level, it may not be possible to obtain sufficient pairings to jointly scale sets unless a method is employed to focus the pairings to enable joint scaling.

The comparison of random pairings and core vs. non-core pairings shows that the latter makes more efficient use of available pairs for the purpose of aligning the two scales. Using the core vs. non-core method, the difference between secondary and primary means was more accurately estimated and the variation of the estimate of this difference was substantially less across simulations, indicating less error in aligning the scales. The gain in efficiency is larger when there is less overlap between the distributions of the two scales and that overlap can be judged based on available information. If the distributions overlap substantially, efficiency is not gained. However, even in this case the advantage still remains that performances can be selected based on fit. Additionally, in more general cases involving three or more sets, two or more separate scales can be efficiently joined through a single core scale, as shown in **Figure 1**. The number of low OUTFIT MNSQ values is larger for the random pairings simulation than in the full joint scaling analysis with core and non-core comparisons. This is likely due to higher level secondary performances being compared favorably against many or all primary performances, in which case many of the residuals are small.

The results of the empirical study indicate reasonably effective scale extension using the core vs. non-core method. The separation index for the scale based on all combined data was high. Fit to the model was not as good as in the simulation study, though reasonable for the applied objectives. In evaluating whether the primary scale could be extended to include secondary performances, a scatterplot was shown of the locations for the secondary (non-core) performances based on the analysis of: (i) the primary and linking set data; and (ii) the secondary data. This is useful to examine whether there is a linear association between the estimates of the core, linking set on the two scales. The scatterplot in **Figure 2** shows a reasonably linear association for the empirical data; the corresponding scatterplot in **Figure 3** shows a clear linear association with a very high correlation for the simulated data.

Once primary and secondary performances were jointly scaled in the empirical context, the secondary performances were qualitatively examined to ascertain whether their positions were defensible relative to the primary performances. These checks were conducted in the form of paired comparisons of secondary and primary, with an emphasis on performances with similar scale locations. The qualitative examination suggested that reasonable alignment of the scales was achieved and that there was not a systematic tendency for secondary performances to be placed too high or too low on the scale relative to the primary performances. In some cases, secondary performances did not appear to be placed well on the scale; however, this is to be expected given the standard errors associated with estimates.

The results of the simulation study showed a close correspondence between: (i) the simulated difference between the primary and secondary mean locations; and (ii) the difference between the mean locations of the estimates of the primary and secondary locations on the joint scale. This confirms that the core vs. non-core comparisons enable extension of the original primary scale with reasonable accuracy when there are a large number of comparisons, using a design of the kind implemented in the empirical study.

First, the method enables utilization of information from the existing scale in designing the scale extension exercise. In the present study, higher level performances were selected from the pre-existing primary scale because the secondary school performances would be compared better more often if lower-level performances had been selected, yielding extreme locations. It is possible to select core performances that have adequate fit also. That is, it is possible to select sets of performances for cross-set comparisons to optimize joint scaling results according to criteria for relative targeting and model fit of performances used.

Second, because the method avoids further within-set comparisons, effort by judges on comparisons is concentrated on comparisons that enable the scales to be equated. Theoretically, the standard errors of estimates in the core and non-core sets will decrease as a result of the addition of core vs. non-core comparisons due to additional Fisher information from additional comparisons. However, theoretically the standard errors of all other estimates will not decrease because there are no further comparisons to provide additional Fisher information. The method is therefore most appropriate where the priority is the efficient use of time available to make comparisons for the equating of scales. Given measurement of a common construct and appropriate targeting and fit, theoretically it is anticipated that a greater number of core vs. non-core comparisons will result in improved alignment of the two scales.

Third, the method potentially provides clearer diagnostic information about the robustness of the joining or equating of the scales than may otherwise be available. The evaluation of the association of locations, shown in Figure 2 for the empirical study and in Figure 3 for the simulation study, are possible due to the design. The objective of the project was to place the performances on a single scale. It is therefore expected that the non-core secondary performances will have the same relative scale locations when derived from comparisons against primary performances as when derived from comparisons against other secondary performances. In the present study, the design enables estimates of the secondary performances solely from comparisons of secondary against primary performances. These were compared with estimates obtained from standard pairwise comparison scaling of the secondary performances to evaluate whether there is a linear association between the independent estimates obtained from the two sets of comparisons.

In addition, for diagnostic purposes, performance-level fit statistics specifically for cross-set comparisons can be obtained to evaluate whether linking set comparisons fit the model adequately. Without the use of a linking set, it is more difficult to focus specifically on diagnostic information related to comparisons that connect the two sets of data.

The comparison of core vs. non-core performances ensures there are cross-set comparisons to enable joint scaling. In a given empirical context, the design and number of comparisons need to be selected to meet accuracy requirements for such applied objectives.

The context of the present study is analogous to vertical equating using an item response model. For equating, core performances need to be reasonably targeted to the non-core performances in terms of the latent ability of students as explained earlier in this article.

Although not the main focus of this article, as discussed above, the generation of core vs. non-core pairs is also applicable where the objective is to obtain scale estimates for performances on a pre-existing scale. That is, the generation of such pairs enables performances to be placed on an existing scale. This opens up the possibility of research into computer adaptive assessment procedures based on: (i) locations on an existing scale; and (ii) estimates of the locations of performances to be placed on the scale obtained after each successive comparison of a performance against a scaled performance. The nature of such an application is virtually identical to computer adaptive testing using IRT estimation. Consequently, practitioners can draw upon relevant literature regarding techniques, algorithms, and so forth as the basis of presenting pairs to judges in the same essential manner that items are presented to students in computer adaptive testing using calibrated item banks.

Limitations and delimitations

The present article aims to illustrate the method and its application in a particular empirical context that enables explanation of key considerations. It is beyond the scope of this article to investigate the number of core and non-core items and number of comparisons required to equate scales. Simulation studies could be used to ascertain the accuracy of equating under different conditions, given combinations of the following parameters: number of core performances; numbers of non-core performance sets; non-core performances per set; and numbers of core vs. non-core comparisons. Although such investigations are beyond the scope of this article, key considerations have been articulated, including the necessity to select core and non-core performances that have overlapping levels of achievement to the extent feasible. Selecting a range of performance levels is also desirable for checking there is a linear relationship as shown in Figures 2, 3.

With respect to the applied objective of the chosen context, the study shows that it is possible to equate primary and secondary persuasive writing scales according to the criteria adopted. The scatterplot showed a reasonably good correlation. Having said this, we consider that it would be ideal to have a higher correlation, above 0.8. The number of comparisons is a key factor affecting the precision of the estimates and, therefore, the highest correlation that can be obtained.

Further research would be needed to examine how generalizable the empirical finding is that primary Writing scales can be extended to include secondary school performances. It is noted, however, that in unpublished studies, primary and secondary English persuasive performances have been jointly scaled as part of the Australian National Assessment Program—Literacy and Numeracy for a number of years. The authors conducted work on these exercises and consider the model fit in such exercises generally good and similar to fit reported in Humphry and McGrane (2015). However, it is beyond the scope of this article to go into further depth about the generalizability of the empirical findings.

In addition to having adequate correlations and person separation indices, the ordering of the performances must validly reflect the latent trait of interest. Attention needs to be given to whether the ordering of the performances is considered to validly reflect the nature of the trait being measured, in terms of the progression of skills evident in performances with increasing scale locations.

Summary and conclusion

This article described and illustrated a method for the joint scaling of two or more sets of performances based on pairwise comparisons and illustrated its application in an empirical context. The article focused on a case in which there were only two sets of performances and subsets of primary and secondary performances were designated core and non-core. This method is applicable where there is an existing scale of student abilities and the objective is to equate one or more new scales onto the existing scale. The method is referred to as a *linking set* scale extension. The method is achieved by selecting a core linking set of performances and by generating core vs. non-core comparisons to equate any number of existing scales.

A simulation study was used to show that the method enables the extension of a scale under conditions similar to those in the empirical study with a larger number of comparisons. This article illustrated the application of the method to a persuasive Writing scale and used this context to summarize key applied considerations. Comparison of random pairings with core/non-core pairings showed the latter is more efficient and that for a given number of pairs, it provided more accurate alignment of the scales and less variation in the alignment across simulations.

The method described in this article can be used to equate two scales provided the scales measure the same latent trait, the two scales are based on responses to tasks of comparable difficulty, and there is sufficient overlap in the level of performances. This method is flexible and efficient, taking advantage of a pre-existing measurement scale to select core performances to extend a scale. A high level of internal reliability was obtained in the empirical study. Assessment of the validity of measurement of the intended construct can be achieved by qualitative examination of the progression of skills and knowledge with increasing scale locations.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the University of Western Australian Education Human Research Ethics Committee. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

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

Funding

This research was partially funded by an Australian Research Council Linkage Project (ARC LP180100046). We acknowledge the funding received for this project from the Australian Research Council, Australian Catholic University, University of Western Australia and Partner Organisations: the Queensland Department of Education, Western Australian School Curriculum and Standards Authority, and Catholic Education Western Australia Limited.

Acknowledgments

We acknowledge the participation of relevant schools in Western Australia in the empirical study reported in this article.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Andrich, D. A. (1978). Relationships Between the Thurstone and Rasch Approaches to Item Scaling. *Appl. Psychol. Meas.* 2, 451–462.

Bradley, R. A., and Terry, M. E. (1952). Rank Analysis of Incomplete Block designs: The Method of Paired Comparisons. *Biometrika* 39, 324–345.

Bramley, T. (2007). "Paired Comparison Methods," in *Techniques for monitoring the comparability of examination standards*, eds P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (London: Qualifications and Curriculum Authority), 246–300.

Bredemeyer, K. (2021a). *pair.generator: Pair performances for assessment using pairwise comparisons. R package version 0.3.2*. Available online at: https://github.com/KenBredemeyer/pair.generator (accessed June 29, 2022).

Bredemeyer, K. (2021b). *PairwiseComparisons: BTL modelling for pairwise comparisons. R package version 0.1.0*. Available online at: https://github.com/KenBredemeyer/PairwiseComparisons (accessed June 29, 2022).

Heldsinger, S. A., and Humphry, S. M. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *Aust. Educ. Res.* 37, 1–19. doi: 10.1007/BF03216919

Heldsinger, S. A., and Humphry, S. M. (2013). Using calibrated exemplars in the teacher-assessment of writing: an empirical study. *Educ. Res.* 55, 219–235. doi: 10.1080/00131881.2013.825159

Humphry, S. M., and Heldsinger, S. A. (2019). A Two-Stage Method for Classroom Assessments of Essay Writing. *J. Educ. Meas.* 56, 505–520. doi: 10.1111/jedm.12223

Humphry, S. M., and Heldsinger, S. A. (2020). A Two-Stage Method for Obtaining Reliable Teacher Assessments of Writing. *Front. Educ.* 56, 505–520. doi: 10.3389/feduc.2020.00006

Humphry, S. M., and McGrane, J. A. (2015). Equating a large-scale writing assessment using pairwise comparisons of performances. *Aust. Educ. Res.* 42, 443–460. doi: 10.1007/s13384-014-0168-6

Luce, R. (1959). *Individual choice behavior*. New York, NY: Wiley.

Pollit, A. (2009). "Abolishing marksism and resuing validity. Cambridge Exam Research," in *A paper for the 35th Annual conference of the International Association for Educational Assessment*, (Brisbane).

Pollit, A. (2012). Comparative Judgement for Assessment. *Int. J. Techol. Des. Educ.* 22, 157–170. doi: 10.1007/s10798-011-9189-x

R Core Team (2021). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., and Sharpe, M. (2008). Rasch Fit statistics and sample size considerations for polytomous data. *Med. Res. Methodol.* 8:33. doi: 10.1186/1471-2288-8-33

Steedle, J. T., and Ferrara, S. (2016). Evaluating Comparative Judgment as an Approach to Essay Scoring. *Appl. Meas. Educ.* 29, 211–223. doi: 10.1080/08957347.2016.1171769

Thurstone, L. L. (1927). A Law of Comparative Judgment. *Psychol. Rev.* 34, 273–286.

Verhavert, S., Bouwer, R., Donche, V., and De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assess. Educ. Princ. Policy Pract.* 26, 541–562. doi: 10.1080/0969594X.2019.1602027

Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., and Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

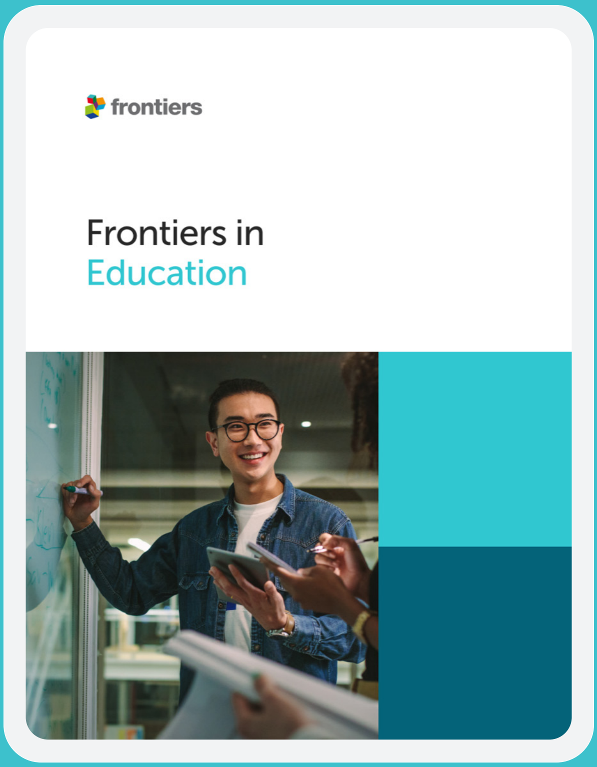
Frontiers in Education

Explores education and its importance for individuals and society

A multidisciplinary journal that explores research-based approaches to education for human development. It focuses on the global challenges and opportunities education faces, ultimately aiming to improve educational outcomes.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Education

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact