frontiers | Research Topics

Exploring classroom assessment practices and teacher decision-making

Edited by
Dennis Alonzo, Chris Davison and Chris Ann Harrison

Published in
Frontiers in Education

ISSN 1664-8714
ISBN 978-2-8325-2408-4
DOI 10.3389/978-2-8325-2408-4

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

Exploring classroom assessment practices and teacher decision-making

Topic editors

Dennis Alonzo — University of New South Wales, Australia
Chris Davison — University of New South Wales, Australia
Chris Ann Harrison — King's College London, United Kingdom

Citation

Alonzo, D., Davison, C., Harrison, C. A., eds. (2023). *Exploring classroom assessment practices and teacher decision-making*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2408-4

Table of contents

frontiers | Frontiers in Education

TYPE Editorial
PUBLISHED 28 April 2023
DOI 10.3389/feduc.2023.1098892

OPEN ACCESS

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

*CORRESPONDENCE
Dennis Alonzo
✉ d.alonzo@unsw.edu.au

RECEIVED 15 November 2022
ACCEPTED 07 April 2023
PUBLISHED 28 April 2023

CITATION
Alonzo D, Davison C and Harrison CA (2023) Editorial: Exploring classroom assessment practices and teacher decision-making. *Front. Educ.* 8:1098892. doi: 10.3389/feduc.2023.1098892

Editorial: Exploring classroom assessment practices and teacher decision-making

Dennis Alonzo[1]*, Chris Davison[1] and Chris Ann Harrison[2]

[1]Shcool of Education, University of New South Wales, Kensington, NSW, Australia, [2]School of Education, Communication & Society, King's College London, London, United Kingdom

KEYWORDS

classroom assessment practices, decision-making, learning, student outcome, teaching

Editorial on the Research Topic
Exploring classroom assessment practices and teacher decision-making

Teaching is a series of decisions teachers make when they plan and deliver activities to help students learn. While some decisions will be taken by the head teacher or district, it is the teachers that are faced with and take the majority of the decisions in the classroom. Evidence of learning is generated as students take part in classroom activities and depend largely on the degree of the teacher's capability to recognize and notice usable information about student learning that they can interpret and use to inform instructional decisions and feedback to students (Bennett, 2011). This process provides actionable information for formative purposes that drive instruction and direct next steps in learning.

Borko et al. (1990) estimated that teachers make around 40–50 decisions in a 1 h lesson. Some of these are planned for within the lesson activities, while others arise during interactions in the classroom. Teachers' decision-making is influenced by their career stages. Experienced teachers call on their recollections of previous lessons to help them make decisions about how to take learning forward as they gauge how their current learners benefit from the lesson activities and use the incoming evidence to decide on next steps. Newer teachers do not have as many experiences to draw upon and will often be meeting student performance on a specific activity for the first time. They also will have less developed assessment knowledge and strategies to be able to respond to the assessment evidence that arises. Peterson and Comeaux (1987) reported that expert and novice teachers differ in the cognitive complexity with which they view classroom events enabling the expert to problem solve more broadly and effectively. Experienced teachers seem more able to focus on the assessment evidence arising from a specific classroom activity and to respond to this in terms of adapting upcoming activities to provide further opportunities for learning rather than taking a narrower view of lesson outcomes which novice teachers tend to do.

Teacher classroom assessment makes up the majority of the assessment activities that a student will experience, and if that assessment is designed to support learning, it can be one of the most powerful interventions to enhance student progress (Black and Wiliam, 1998; Hattie, 2008; Alonzo, 2020). When it comes to the concept of using assessment to support learning, many terms are used interchangeably to refer to similar assessment practices and procedures, including terms such as classroom-based assessment, formative assessment, assessment for learning, and, more recently, learning-oriented assessment. These terminologies all refer to pedagogically-linked assessment approaches that require embedding any assessment in learning and teaching processes to promote student learning.

The central role of teacher assessment practices in improving student learning has gained significant attention and has been extensively researched.

In parallel with this increasing focus on teacher assessment practices is a growing interest in the factors and processes involved in teacher decision-making because of their critical importance in improving learning and teaching (Bianco, 2010; Mandinach and Schildkamp, 2021a). Teacher decision-making is seen as an integral component of teacher assessment practice (Mandinach and Schildkamp, 2021b; Beswick et al., 2022). However, despite the strong pressure for teachers to use assessment information to inform instructional decisions, major drawbacks are reported in the literature. These include the capacity of teachers to translate information into insights (Datnow and Hubbard, 2015), the amount of time and onerous preparation needed (Datnow et al., 2021), equity concerns (Dodman et al., 2021), access and availability of various kinds of data (Kallemeyn, 2014), and data system design and construction (Drake, 2021). There are also decisions as to which type of information can best support teacher decision-making with differing understandings of what constitutes assessment information, some considering only the system-level data generated through standardized testing to be rigorous enough to provide insights to inform teacher practice. However, data for teacher decision-making can include "student achievement (from qualitative teacher records to high-stake tests), socio-demographic and contextual information about schools, teachers and students, and non-cognitive characteristics of students, teachers and school leaders (Beswick et al., 2022, p. 2)." Apart from these issues, teacher decision-making is also marred by competing evidence, with some studies showing no impact on student learning (Reeves and Burt, 2006; Staman et al., 2017). Hence, there is a need to gather more evidence to address these gaps identified in the literature.

Our Research Topic draws on current research adding to the growing evidence of the importance of teacher assessment practices and decision-making. There are 11 papers included in this Research Topic with diverse aims.

At the classroom level, Earle's paper explores formative decision-making and the subsequent actions taken by teachers to inform learning and teaching. Her study reports that teacher decision-making informed by formative assessment data leads to immediate or future changes in learning and teaching activities. Näsström et al. describe one teacher's formative assessment practice and the requirements for effective teacher decision-making. Their study found that students in the intervention teacher's class increased their controlled and autonomous forms of motivation as well as their engagement in learning activities. In addition, Cowie et al. demonstrate how to use a Data Conversation Protocol to analyze and act on mathematics assessment data generated through a standardized assessment tool. The Conversation Protocol helps teachers to slow down the process of considering, interpreting and making a judgement about their students' understanding. They also found that students responded positively to teachers' data informed small group teaching, gaining in understanding and confidence. Further, Monteiro et al. examine how teachers and students view assessment and how teachers assess their students' learning, how teachers assess their students' learning, and the similarities and disparities that occur when students' and teachers' conceptions and teachers' practices of assessment are compared. Their results show that teachers' conceptions of assessment contradict their actual assessment practices. In addition, their study shows that students' conceptions of assessment are constructed from their classroom assessment experiences.

Three studies offer a broader understanding of teacher assessment and decision making skills. Gu offers an argument-based framework for validating formative assessment in the classroom. He offered an operational definition of formative assessment and classroom-based formative assessment. He argues that a clear operationalization is the starting point for researchers and teachers alike to examine the validity and effectiveness of the formative assessment construct. van der Steen et al. create a set of design principles to support teachers in designing formative assessment plans informing formative decision-making. Based on expert interviews expert interviews and subsequent evaluation of future users, there are eight suggested design principles that can be used and validated in educational practice. Phung and Michell report on the nature and dynamics of teacher decision-making, and conceptualized assessment decision-making pathways. They propose three assessment decision-making pathways which provide a new lens for understanding differences in teachers' final assessment judgements of student oral language performances and their relative trustworthiness.

Three studies focus on pre-service teachers. Schnitzler et al. investigate how student teachers with high and low judgment accuracy differ with regard to their eye movements as a behavioral and utilization of student cues as a cognitive activity. Their findings highlight the power of behavioral and cognitive activities in judgment processes for explaining teacher performance of judgment accuracy. Kron et al.'s study focuses on pre-service mathematics teachers' selection of tasks during one-to-one diagnostic interviews in a live simulation. The results highlight that pre-service teachers require further support to effectively attend to diagnostic task potential. Oo et al. report on the results of a study of the process of preservice teachers' decision-making in assessment practices in Myanmar. They have demonstrated how beliefs and values shape pre-service teachers' assessment practices. Lastly, Alonzo et al. report on a case study of a school in building an assessment culture with a strong focus on using a range of data for teacher decision making. Using the lens of activity theory, they have identified structural, organizational, social, and behavioral factors that contribute to the success of the program.

Despite the range of Research Topic reported in this special issue and extant literature, a continuous exploration of this critical enquiry is required to provide a more nuanced understanding of teacher decision-making skills. As argued above, effective teaching happens when teachers are engaged in ongoing decision-making. Thus, it is important that we further advance the theorisation of this construct to support teachers to improve their decision-making skills, making their practices more trustworthy.

Author contributions

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

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

Alonzo, D. (2020). "Teacher education and professional development in industry 4.0. The case for building a strong assessment literacy," in *Teacher Education and Professional Development in Industry 4.0. 4th International Conference on Teacher Education and Professional Development (InCoTEPD 2019)*, eds J. Ashadi, A. Priyana, A. Basikin, A. Triastuti, and N. Putor (London: Taylor & Francis Group), 3–10.

Bennett, R. (2011). Formative assessment: A critical review. *Assess. Educ.* 18, 5–25. doi: 10.1080/0969594X.2010.513678

Beswick, K., Alonzo, D., and Lee, J. (2022). *Data Literacy for Student Outcomes: Supporting Principals and Teachers to Use Data for Evidence-Informed Decision-Making*. Kensington, NSW: School of Education, University of New South Wales.

Bianco, S. D. (2010). Improving student outcomes: Data-driven instruction and fidelity of implementation in a response to intervention (RTI) model. *Teaching Except. Child. Plus* 6, 1–13. Available online at: http://escholarship.bc.edu/education/tecplus/vol6/iss5/art1

Black, P., and Wiliam, D. (1998). Assessment and classroom learning'. *Assess. Educ.* 5, 7–74. doi: 10.1080/0969595980050102

Borko, H., Livingston, C., and Shavelson, R. J. (1990). Teachers' thinking about instruction. *Remedial Spec. Educ.* 11, 40–49. doi: 10.1177/074193259001100609

Datnow, A., and Hubbard, L. (2015). Teachers' use of assessment data to inform instruction: Lessons from the past and prospects for the future. *Teachers Coll. Record* 117, 1–45. doi: 10.1177/016146811511700408

Datnow, A., Lockton, M., and Weddle, H. (2021). Capacity building to bridge data use and instructional improvement through evidence on student thinking. *Stud. Educ. Eval.* 69, 100869. doi: 10.1016/j.stueduc.2020.100869

Dodman, S. L., Swalwell, K., DeMulder, E. K., View, J. L., and Stribling, S. M. (2021). Critical data-driven decision making: A conceptual model of data use for equity. *Teach. Teacher Educ.* 99, 103272. doi: 10.1016/j.tate.2020.103272

Drake, T. A. (2021). "We have all the data in one place": Examining principals' use of a data warehouse during an academic school year. *NASSP Bullet.* 105, 84–110. doi: 10.1177/01926365211015311

Hattie, J. (2008). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Hoboken, NJ: Routledge.

Kallemeyn, L. M. (2014). School-level organisational routines for learning: Supporting data use. *J. Educ. Admin.* 52, 529–548. doi: 10.1108/JEA-02-2013-0025

Mandinach, E. B., and Schildkamp, K. (2021a). The complexity of data-based decision making: An introduction to the special issue. *Stud. Educ. Eval.* 69, 100906. doi: 10.1016/j.stueduc.2020.100906

Mandinach, E. B., and Schildkamp, K. (2021b). Misconceptions about data-based decision making in education: An exploration of the literature. *Stud. Educ. Eval.* 69, 100842. doi: 10.1016/j.stueduc.2020.100842

Peterson, P. L., and Comeaux, M. A. (1987). Teachers' schemata for classroom events: The mental scaffolding of teachers' thinking during classroom instruction. *Teach. Teacher Educ.* 3, 319–331 doi: 10.1016/0742-051X(87)90024-2

Reeves, P. L., and Burt, W. L. (2006). Challenges in data-based decision-making: voices from principals. *Educ. Horizons* 84, 65. Available online at: https://www.jstor.org/stable/42925967

Staman, L., Timmermans, A. C., and Visscher, A. J. (2017). Effects of a data-based decision making intervention on student achievement. *Stud. Educ. Eval.* 55, 58–67. doi: 10.1016/j.stueduc.2017.07.002

frontiers in Education

ORIGINAL RESEARCH
published: 10 December 2020
doi: 10.3389/feduc.2020.602470

Connecting Judgment Process and Accuracy of Student Teachers: Differences in Observation and Student Engagement Cues to Assess Student Characteristics

Katharina Schnitzler[1]*, *Doris Holzberger*[2] *and Tina Seidel*[1]

[1] TUM School of Education, Chair for Educational Psychology, Technical University of Munich, Munich, Germany, [2] TUM School of Education, Centre for International Student Assessment, Technical University of Munich, Munich, Germany

OPEN ACCESS

Edited by:
Chris Davison,
University of New South Wales, Australia

Reviewed by:
Peter Nyström,
University of Gothenburg, Sweden
Hui Yong Tay,
Nanyang Technological University, Singapore

***Correspondence:**
Katharina Schnitzler
Katharina.Schnitzler@tum.de

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

Received: 03 September 2020
Accepted: 13 November 2020
Published: 10 December 2020

Citation:
Schnitzler K, Holzberger D and Seidel T (2020) Connecting Judgment Process and Accuracy of Student Teachers: Differences in Observation and Student Engagement Cues to Assess Student Characteristics. Front. Educ. 5:602470. doi: 10.3389/feduc.2020.602470

Teachers' ability to assess student cognitive and motivational-affective characteristics is a requirement to support individual students with adaptive teaching. However, teachers have difficulty in assessing the diversity among their students in terms of the intra-individual combinations of these characteristics in student profiles. Reasons for this challenge are assumed to lie in the behavioral and cognitive activities behind judgment processes. Particularly, the observation and utilization of diagnostic student cues, such as student engagement, might be an important factor. Hence, we investigated how student teachers with high and low judgment accuracy differ with regard to their eye movements as a behavioral and utilization of student cues as a cognitive activity. Forty-three participating student teachers observed a video vignette showing parts of a mathematics lesson to assess student characteristics of five target students, and reported which cues they used to form their judgment. Meanwhile, eye movements were tracked. Student teachers showed substantial diversity in their judgment accuracy. Those with a high judgment accuracy showed slight tendencies toward a more "experienced" pattern of eye movements with a higher number of fixations and shorter average fixation duration. Although all participants favored diagnostic student cues for their assessments, an epistemic network analysis indicated that student teachers with a high judgment accuracy utilized combinations of diagnostic student cues that clearly pointed to specific student profiles. Those with a low judgment accuracy had difficulty using distinct combinations of diagnostic cues. Findings highlight the power of behavioral and cognitive activities in judgment processes for explaining teacher performance of judgment accuracy.

Keywords: judgment accuracy, judgment process, lens model, student cue utilization, student engagement, student profiles, epistemic network analysis, eye tracking

INTRODUCTION

Teacher assessment skills are an essential component of professional competence (Baumert and Kunter, 2006, 2013; Binder et al., 2018). In their daily professional lives, teachers are required to continuously make educational decisions when assigning grades, planning lessons, adapting their teaching, and providing feedback. To effectively make informed decisions, teachers must constantly monitor their students' learning-relevant cognitive (e.g., cognitive abilities or knowledge) and motivational-affective characteristics (e.g., academic self-concept or interest) and the specific combination of these characteristics within individual students (Corno, 2008; Herppich et al., 2018; Heitzmann et al., 2019; Loibl et al., 2020). Some students may, for example, possess high cognitive characteristics combined with low motivational-affective characteristics, indicating underestimation of their abilities. Other students are aware of their abilities, and hold high cognitive and motivational-affective characteristics (Seidel, 2006). Students with such varying profiles differ in how they engage with, achieve in, and experience their learning environment (Seidel, 2006; Lau and Roeser, 2008; Jurik et al., 2013, 2014), and their positive educational development depends on tailored teacher instruction (Huber et al., 2015). Therefore, it is alarming to know that teachers are struggling to accurately assess the diversity of cognitive and motivational-affective characteristics among their students (Huber and Seidel, 2018; Südkamp et al., 2018). To assess student profiles accurately, teachers are required to observe their students for relevant cues, such as the intensity and content of engagement, and use combinations of such cues to infer underlying combinations of student characteristics (Cooksey et al., 1986; Nestler and Back, 2013; Thiede et al., 2015). Yet systematic linkages of judgment processes and judgment accuracy are rather rare (Karst and Bonefeld, 2020), and are still unclear in terms of how they are used to assess student profiles (Praetorius et al., 2017; Huber and Seidel, 2018). Thus, the aim of the present study is to explore such connections and contribute to existing research in the field by providing detailed insights into the process of observing and assessing student profiles. Therefore, we investigate how well teachers are able to accurately judge various student profiles. Moreover, we link this judgment accuracy to two factors: eye movements (as a measure of the behavioral activity of observing students) and utilization of student cues (as a measure of the cognitive activity) behind judgment processes.

Student Characteristic Profiles as Targets for Teacher Assessment

Since the seminal work of Snow (1989), cognitive and motivational-affective student characteristics are seen as fundamental determinants of learning and achievement. Robust empirical studies with large representative samples and meta-analyses have shown that cognitive abilities (Deary et al., 2007; Roth et al., 2015), pre-achievement (Steinmayr and Spinath, 2009), academic self-concept as students' perception of their subject-specific abilities (Shavelson et al., 1976; Valentine et al., 2004; Steinmayr and Spinath, 2009; Huang, 2011; Marsh and Martin, 2011), and subject interest (Schiefele et al., 1992; Jansen et al., 2016) are among the most decisive student characteristics for educational outcomes.

Consistent and Inconsistent Combinations of Cognitive and Motivational-Affective Characteristics

A strand of research has begun to investigate the complex and interrelated influences that cognitive and motivational-affective characteristics might have on student learning. Therefore, researchers have followed a person-centered approach to examine the intra-individual interplay of student characteristics for the purposes of identifying which combinations of cognitive and motivational-affective characteristics are predominant among students (Seidel, 2006; Lau and Roeser, 2008; Linnenbrink-Garcia et al., 2012; Seidel et al., 2016; Südkamp et al., 2018).

Seidel (2006), for example, used a latent-cluster analysis to identify homogenous subgroups of students that are distinct from one another, in that each subgroup showed a unique pattern of cognitive characteristics—cognitive abilities and pre-knowledge—combined with subject interest and academic self-concept as motivational-affective characteristics. Five so-called student profiles were identified. Two of these profiles can be seen as "consistent," in that they are assigned to individuals who displayed either low or high levels of cognitive *and* motivational-affective characteristics: First, "strong" students were very likely to show high values for all characteristics. Second, students who were likely to show low values for all characteristics and were labeled as "struggling." The remaining three profiles are considered to be "inconsistent," since the interplay of cognitive and motivational-affective characteristics within individuals to whom these profiles are assigned are either opposing or non-uniform: "Overestimating" students showed relatively low values for cognitive characteristics but were likely to report high subject interest and positive self-concept. Hence, these students might overestimate their abilities. "Underestimating" students displayed an opposite pattern in which high cognitive abilities were combined with low interest and low self-concept. These students seemed to underestimate their abilities. Finally, "uninterested" students stood out due to their high cognitive abilities and particularly low subject interest. Altogether, 57% of the students investigated by Seidel (2006) belonged to inconsistent profiles.

Looking at student diversity from the viewpoint of student characteristic profiles is meaningful, since other studies have repeatedly found mixtures of consistent and inconsistent profiles (Lau and Roeser, 2008; Linnenbrink-Garcia et al., 2012; Seidel et al., 2016; Südkamp et al., 2018). In all of these studies, there was a significant proportion of students that shared inconsistent profiles, ranging in studies from 10% (Südkamp et al., 2018) to more than half of investigated students (Seidel, 2006; Linnenbrink-Garcia et al., 2012). This seems to be a generalizable finding, since the reviewed studies are spread across different subjects including physics, science, biology, mathematics, and language arts, and addressed different cognitive (e.g., cognitive abilities, pre-knowledge, grades) and motivational-affective characteristics (e.g., academic self-concept, learning motivation, anxiety, task-value; Seidel, 2006; Lau and Roeser,

2008; Linnenbrink-Garcia et al., 2012; Seidel et al., 2016; Südkamp et al., 2018). For teachers, these inconsistent profiles are meaningful and quite likely to be present in every classroom. One way in which specific differences between the profiles become apparent to teachers is through student engagement as a central component of student learning.

Student Engagement Reflects Student Characteristic Profiles

Relationships exist between student characteristic profiles and engagement in learning activities as a precondition for achievement. These relationships have been examined in research from two perspectives: first using students' own reports of classroom experience as an antecedent of their engagement; and second by proximal assessments of student engagement through self-reports and video observation. Students with high motivational-affective characteristics seem to perceive their learning environment as particularly supportive and experience high-quality teaching (Seidel, 2006; Lau and Roeser, 2008). Along with these positive perceptions, students with high motivational-affective characteristics are also more frequently cognitively and behaviorally engaged in learning activities independent of the level of their cognitive characteristics. They report high levels of elaborating and organizing information (Jurik et al., 2014), attention, and participation in learning activities and classroom talk (Lau and Roeser, 2008), and show especially high numbers of verbal interactions with teachers (Jurik et al., 2013). In contrast, students with low motivational-affective characteristics often perceive their learning environment in a negative way (Seidel, 2006; Lau and Roeser, 2008) and suffer from low engagement (Lau and Roeser, 2008; Jurik et al., 2013, 2014). These differences in engagement result in differential effects on student learning and achievement (Lau and Roeser, 2008; Linnenbrink-Garcia et al., 2012).

Diversity in terms of characteristic profiles shapes students' classroom experiences and engagement, and in turn, educational achievement. Therefore, it is argued that teachers need to be aware of these prototypical profiles if they want to effectively support student learning (Huber and Seidel, 2018). Moreover, to make appropriate educational decisions and take effective actions, teachers must also be able to correctly assess the complex combinations of cognitive and motivational-affective characteristics of individual students (Praetorius et al., 2017).

Teacher Judgment Accuracy of Student Characteristic Profiles

To date, research has focused mainly on how accurately teachers judge individual student characteristics, and less on how they assess the interplay of characteristics through student profiles. It is important to note, however, that the ability to achieve the former is a necessary precondition to performing the latter (Südkamp et al., 2018). According to meta-analyses, teachers make relatively accurate judgments of students' cognitive abilities (Machts et al., 2016) and achievement (Südkamp et al., 2012). The few studies that deal with judgment accuracy of motivational-affective characteristics showed that teachers are only somewhat able to accurately assess students' self-concept (Spinath, 2005; Praetorius et al., 2013; Urhahne and Zhu, 2015) and interest (Karing, 2009). Hence, teachers seem to have more difficulties assessing student motivational-affective characteristics than cognitive characteristics (Kaiser et al., 2013; Praetorius et al., 2017). Moreover, that teachers intermingle single student characteristics, for example, when prompted to assess achievement and motivation separately, indicates that they tend to perceive students holistically (Kaiser et al., 2013). As a result of this phenomenon, it is particularly important to focus on how teachers assess student profiles that combine cognitive and motivational-affective characteristics.

So far, few studies have addressed this issue. Two studies have shown that teachers tend to underestimate the extent of inconsistent profiles among their students (Huber and Seidel, 2018; Südkamp et al., 2018). Teachers seem to assume that cognitive and motivational-affective characteristics typically go hand in hand, and subsequently categorize their students simply as being average, below-average, or above-average (Huber and Seidel, 2018; Südkamp et al., 2018). However, when teachers are explicitly asked to assign students to consistent and inconsistent profiles—when the degree of inconsistency itself is not in question—it was shown that experienced teachers are more accurate in assessing inconsistent student profiles than student teachers, although a considerable amount of variance was apparent among experienced and student teachers alike (Seidel et al., 2020). These differences in judgment accuracy might originate from differences in the preceding judgment process (Loibl et al., 2020). Therefore, to understand why some teachers achieve high judgment accuracy when assessing student profiles while others fail to do so, it is necessary to investigate in more detail the processes of judgment formation. As research has focused predominantly on teacher judgment accuracy, far less is known about the cognitive and behavioral activities that drive the judgment process itself, especially when it comes to the connection between judgment process and judgment accuracy (Herppich et al., 2018; Karst and Bonefeld, 2020; Loibl et al., 2020).

Teachers' Process for Judging Student Characteristic Profiles

Judgment processes comprise behavioral and cognitive activities (Loibl et al., 2020). Since teaching is a vision-intense profession in which it is important to gain information by monitoring what is happening in the classroom (Carter et al., 1988; Gegenfurtner, 2020), the observation of students to gain information (i.e., behavioral activity) and the interpretation of this information to make decisions (i.e., cognitive activity) are likely to be relevant activities of judgment processes. The ability to succeed in these activities is recognized as a central component of a teacher's professional competence (Blömeke et al., 2015; Santagata and Yeh, 2016), and is often labeled as professional vision (Goodwin, 1994; van Es and Sherin, 2002, 2008; Sherin and van Es, 2009; Seidel and Stürmer, 2014). In psychological research, the so-called lens model, which is based on Brunswik's (1956) paradigm that humans observe and interpret information cues to make sense of their ambiguous environment, systematizes this idea.

The model is regularly considered in other fields involving judgment formation such as social sciences, business science, and medicine (Cooksey et al., 1986; Funder, 1995, 2012; Kaufmann et al., 2013; Kuncel et al., 2013). It also receives attention in the educational field (Cooksey et al., 1986, 2007; Marksteiner et al., 2012; Thiede et al., 2015; Praetorius et al., 2017).

According to the lens model, teachers are required to *observe* and *utilize*—that is combine and interpret—several student behaviors (i.e., student cues) to inform themselves about student characteristics (Cooksey et al., 1986; Nestler and Back, 2013; Thiede et al., 2015). Therefore, the manifestation of student characteristics in specific observable student cues is a precondition to judgment. Such student cues are referred to as "diagnostic" (Funder, 1995; Thiede et al., 2015) or "ecologically valid" (Cooksey et al., 1986; Nestler and Back, 2013; Back and Nestler, 2016). In other words, accurate judgments depend on the observation and use of diagnostic student cues (Nestler and Back, 2013; Förster and Böhmer, 2017). To do so, teachers require a professional knowledge base, which allows them to connect student cues to underlying student characteristics (Funder, 1995, 2012; Meschede et al., 2017). In this sense, successful judgment processes represent an applied form of professional knowledge of teachers (Jacobs et al., 2010; Stürmer et al., 2013; Kersting et al., 2016; Lachner et al., 2016). Therefore, the lens model provides a suitable framework for the investigation of teachers' behavioral and cognitive activities in the process of accurately judging latent, and not directly observable, student profiles (Nestler and Back, 2013; Förster and Böhmer, 2017; Praetorius et al., 2017; Loibl et al., 2020).

Observation of Students as a Behavioral Activity in the Judgment Process

Eye movements are an indicator for teacher observation behavior (Gegenfurtner, 2020; Loibl et al., 2020), and fall into one of two categories: saccades and fixations. Saccades are fast movements in which the eye is turned for the purposes of bringing objects of interest in front of the fovea so that they can be seen sharply. Fixations are moments when the eye is relatively still and visual information is processed (Holmqvist et al., 2011; Krauzlis et al., 2017). The location of fixations, that is the object on which one fixates, as well as the number and duration of fixations on an object, are driven by top-down and bottom-up processes through declarative knowledge (e.g., knowing where to look for relevant information) and saliency of situational features that attract attention (e.g., student movements such as hand raising behavior or visual features as bright colored clothing), respectively (DeAngelus and Pelz, 2009; Schütz et al., 2011; Gegenfurtner, 2020).

Eye tracking—which measures where one is looking—is a relatively new method in educational science (Jarodzka et al., 2017). Nevertheless, some studies have already provided initial evidence concerning teachers' observation behavior. This evidence comes primarily in the form of comparisons between experienced and student teachers in the context of professional vision (Stürmer et al., 2017; Wyss et al., 2020), classroom management (van den Bogert et al., 2014; Cortina et al., 2015; Wolff et al., 2016), and teacher-student interactions (McIntyre et al., 2017, 2019; McIntyre and Foulsham, 2018; Haataja et al., 2019, 2020; Seidel et al., 2020). Overall, in comparison with student teachers, experienced teachers seem to show a more knowledge-driven pattern of eye movement, which represents selective viewing and fast information processing. Experienced teachers also focus more on areas that are rich in information and pay more attention to students than to other things in the classroom. Moreover, experienced teachers continuously monitor the classroom as a whole even if they are in the process of recognizing relevant events or interacting with individual students (van den Bogert et al., 2014; Cortina et al., 2015; Wolff et al., 2016; McIntyre et al., 2017, 2019; McIntyre and Foulsham, 2018; Wyss et al., 2020). Therefore, experienced teachers show a pattern of monitoring relevant areas with more fixations but shorter fixation durations (van den Bogert et al., 2014; Seidel et al., 2020), similar to experts in other domains (Gegenfurtner et al., 2011). So far, only one study has connected teachers' judgment accuracy with eye movements. Hörstermann et al. (2017) investigated primacy effects concerning the location of information cues in case vignettes for students' social background and performance. It was found that student teachers paid most attention to the information presented at the top left of the case vignettes. The type of information presented in this location, whether related to students' social background or performance, did not bias the accuracy of decisions concerning school track. The available research on teacher eye movement suggests that it can be an appropriate method for gaining additional information about teacher judgment processes. Therefore, it can be used to study, for example, whether teachers, who formed accurate student judgments as the result of a judgment process, also showed an "experienced" pattern of eye movement such as faster information processing, indicating a top-down driven process of advanced knowledge organization.

Utilization of Student Cues as a Cognitive Activity in the Judgment Process

In terms of accurately judging student profiles, teachers are required to assess student characteristics and their intra-individual consistency. In this case, several diagnostic student cues that point toward the level of cognitive and motivational-affective characteristics need to be used in combination. In particular, the *intensity* and *content* of engagement can be considered as relevant diagnostic student cues (see section Student engagement reflects student characteristic profiles for the connection of student profiles and student engagement). With *intensity* of student engagement, we refer to its level of presence. With regard to behavioral aspects, for example, rare hand-raising behavior represents lower intensity of engagement, while frequent hand-raising behavior represents higher intensity of engagement. By *content* of student engagement we refer to the level of knowledge and understanding that becomes apparent through student engagement. For example, when engaging verbally in teacher-student interactions, correctness of an answer or use of technical language represent the content of student engagement. To distinguish, for example, students with strong and overestimating profiles (Seidel, 2006) frequent hand-raising behavior (intensity of engagement) might be a diagnostic cue

for a high level of self-concept (Böheim et al., 2020; Schnitzler et al., 2020), whereas incorrect answers (content of engagement) point toward low knowledge, and correct answers (content of engagement) indicate high knowledge (Thiede et al., 2015). Consequently, only if teachers utilize combinations of diagnostic student cues containing information about cognitive and motivational-affective characteristics, they may infer the correct student profile. Otherwise, profiles that share a similar level of cognitive (e.g., strong and underestimating) or motivational-affective characteristics (e.g., strong and overestimating) might be interchanged.

Empirical findings regarding the question of how well teachers are able to utilize diagnostic student cues are limited. In general, experienced teachers are much better able to interpret relevant classroom events than student teachers and beginning teachers (Sabers et al., 1991; Berliner, 2001; Star and Strickland, 2008; Meschede et al., 2017; Kim and Klassen, 2018; Keppens et al., 2019). This is due to an encapsulated knowledge structure along cognitive schemata which results from the integration of practical experiences with declarative knowledge (see for a current review on expertise development in domains that focus on diagnosing Boshuizen et al., 2020), allowing for fast information processing (Carter et al., 1988; Berliner, 2001; Kersting et al., 2016; Lachner et al., 2016; Kim and Klassen, 2018). However, differences in the abilities to interpret classroom events already appear among student teachers who had only limited opportunities to engage in teaching practice (Stürmer et al., 2016). When it comes to the explicit consideration of judgment processes, teachers seem to consider student background characteristics such as gender, ethnicity, immigration status, and socioeconomic status (SES) to assess students' cognitive and motivational-affective characteristics (Meissel et al., 2017; Praetorius et al., 2017; Garcia et al., 2019; Brandmiller et al., 2020). Moreover, teachers seem to rely on these rather unimportant and misleading student cues especially when they experience low accountability for their decisions (Glock et al., 2012; Krolak-Schwerdt et al., 2013). Additionally, student teachers tend to utilize as many student cues as available, irrespective of whether they are diagnostic or unimportant, while experienced teachers seem to do so only if available cues are inconsistent (Glock et al., 2013; Böhmer et al., 2015, 2017).

Only a small number of studies have investigated teachers' use of cues in connection with their judgment accuracy. For example, beginning teachers seem to be aware of diagnostic cues for detecting whether someone is telling the truth or lying when observing videos. However, cues were utilized in a way that led to inaccurate judgments (Marksteiner et al., 2012). Another study investigated the effect of the availability of different cues (only students' names; students' name and answers on practice tasks; and only students' answers on practice tasks) on the type of information used to assess students' performance on a set of mathematical tasks. Teachers were most accurate in assessing low performance if they knew only students' answers on the practice tasks because under this condition they used more answer-related, diagnostic information (Oudman et al., 2018). When analyzing which cues experienced and student teachers utilize to assess student profiles from video observation, prior findings indicate that these two groups do not differ in the number of student cues utilized. However, experts seem to use a broader range of cues, while student teachers tended to focus more on rather salient student cues, such as frequency of hand-raising (Seidel et al., 2020).

Based on the studies summarized above, it still remains quite unclear which student cues, diagnostic or unimportant, and which student cue combinations are utilized by teachers in everyday teaching to assess cognitive and motivational characteristics, not to mention their combination in student profiles (Glock et al., 2013; Praetorius et al., 2017; Huber and Seidel, 2018; Brandmiller et al., 2020). Furthermore, a link between judgment accuracy and judgment process remains to be established. For example, how do teachers, who succeed in assessing student profiles, differ from those who have difficulty doing so? Do they utilize student cues in a different way?

The Present Study

Against this background, the present study aimed to expand research on the connection between judgment accuracy and judgment process. Furthermore, we considered student diversity in terms of previously identified consistent and inconsistent student profiles as identified by Seidel (2006), and took into account observation of students and utilization of student cues as behavioral and cognitive activities, respectively, as drivers of judgment processes. In addition, we focused on student teachers and the differences previously determined to exist among this group. We addressed the following three research questions:

RQ1 concerning judgment accuracy:

a) How accurately can student teachers judge student profiles?
b) How does student teachers' judgment accuracy differ across student profiles?
c) Which student profiles do student teachers interchange predominantly?

Considering previous findings, we assumed that some student teachers would display high judgment accuracy while others would struggle to assign student characteristic profiles. We expected student teachers to assess consistent profiles with a high judgment accuracy and inconsistent profiles with a lower accuracy due to previous findings, which reported that teachers systematically underestimate the level of diversity among their students. Moreover, we assumed that they would interchange profiles that share the same level of cognitive characteristics but that differ in their motivational-affective characteristics, since teachers were previously found to be better able to assess cognitive characteristics with a higher accuracy.

To deepen our understanding on the interdependence of judgment accuracy and judgment processes we considered two process indicators—behavioral and cognitive activities.

RQ2 concerning observation of students as a behavioral activity:

Across different student profiles, what differences indicate high and low judgment accuracy in student teachers' eye movements

a) with regard to the number of fixations?
b) with regard to the average fixation duration?

We expected that student teachers with a higher judgment accuracy would show a pattern of a higher number of fixations with shorter average duration than those with low judgment accuracy.

RQ3 concerning utilization of student cues as a cognitive activity:

a) Which student cues do student teachers utilize to assess student cognitive and motivational-affective characteristics (student profiles)?
b) What combinations of student cues do student teachers with high and low judgment accuracy use to assess student cognitive and motivational-affective characteristics (student profiles)?

This research question was explorative, given its novelty. Nevertheless, we expected that student teachers with a high judgment accuracy would utilize combinations of diagnostic student cues that reflect both student cognitive and motivational-affective characteristics and point particularly to different student profiles.

METHODS

Participants

Forty-three student teachers ($M_{Age} = 21.59$; $SD = 1.60$; 62.8% female) participated in our study during their fourth semester of a bachelor's teacher training program at the Technical University of Munich. All participants enrolled in a program to become teachers in German high-track secondary schools for science and/or mathematics. We invited student teachers to participate in the study during one of their pedagogical courses. Participants received a 20 Euro voucher.

Procedure

The present study was conducted in line with the Ethical Principles of Psychologists and Code of Conduct of the American Psychological Association from 2017 (APA American Psychological Association, 2017). Participants have been assured that their data will be used in accordance with the data protection guidelines and analyzed for scientific purposes only. They gave informed consent before participation.

Data collection took place in the university laboratory. At the study's outset, participants were familiarized with previously identified student profiles (Seidel, 2006): individual student cognitive and motivational-affective characteristics, as well as their interplay in the form of profiles (strong, struggling, overestimating, underestimating, and uninterested), were illustrated. This included descriptions of each student characteristic avoiding student cues that might be observable in classrooms. Next, to make participants familiar with the classroom environment and the lesson topic, they watched a short video trailer (2:30 min) showing the class in question.

Participants then encountered the assessment situation of this study. We instructed participants to assess the profiles of five target students shown in an 11-min video. Each target student was marked with a random letter (B, E, K, P, T), so that participants were always aware of them (**Figure 1**). While participants watched the video, eye tracking was conducted. Finally, participants were asked to assign a profile to each target student. In doing so, they were also asked to voluntarily indicate, in an open answer format, the student cues they had utilized for their assessment. Each student profile could only be assigned once.

Video Stimulus

The video presented during the assessment stemmed from a previous study on teacher–student interactions in classrooms (Seidel et al., 2016), and showed an eighth grade mathematics lesson in a German high-track secondary school in which a new topic was being introduced. The video consisted of two segments. The first segment showed a teacher describing a task to be accomplished in a subsequent individual work phase. The second segment showed students sharing their results with the teacher after the individual work phase. The individual work itself was not included in the video stimulus; instead, a short text informed participants when students were working on the individual task. Both segments comprised several scenes of students listening to their teacher's lecture, and of students interacting with their teacher in a classroom dialogue. Details about the video segments and scenes are provided in **Figure 1**.

Each target student represented one student profile (strong, struggling, overestimating, underestimating, and uninterested). The students' profiles were empirically determined in a previous video study from which the video stemmed (Seidel et al., 2016). We carefully selected the five target students to best represent a particular student profile with regard to observable student cues. To achieve this, three researchers involved in the present study ranked the students independently in terms of their representation of the identified student profile.

Apparatus

We recorded participants' eye movements with the SMI RED 500 binocular remote eye tracker using Experiment Center software version 3.7 (SMI, 2017b) on a 22-inch display monitor and at a sampling frequency of 500 Hz. Eye tracking conditions were standardized for all participants. Light conditions were kept stable by closing the window blinds and using ceiling lighting. Participants were positioned 65 cm in front of the eye tracker. To increase the precision of eye tracking, a height-adjustable table was used in combination with a chin rest to ensure that the equipment was adjusted to each individual participant, and prevented them from performing strong (head) movements (Nyström et al., 2013). Moreover, before beginning eye tracking, a 9-point automatic calibration followed by a validation was implemented to ensure data quality.

Measurements

Judgment Accuracy

To measure judgment accuracy, participants received one accuracy point if the assigned profile matched the underlying data-driven profile, or no point (wrong profile assigned) for each

Segment	Scene	Duration in minutes	Teacher visibility	Eye-Tracking	Description
1	Interacting	3:20	X	✓	Teacher introduces topic in a question-answer style. Various students raise their hands and provide short answers.
	Listening	0:30	X	✓	Teacher explains the upcoming task. Students listen.
			Cut-out Individual Work Scene		
2	Interacting	1:30	X	✓	Teacher asks for solution of the task. A few students raise their hands and share their ideas.
	Listening	1:25	X	✓	Teacher explains a part of the solution in detail. Students listen.
	Interacting	2:00	X	✓	Teacher continues to ask for solution of the task. A few students raise their hands and share their ideas.
	Individual Work	2:00	✓	X	Students correct their own solution.

FIGURE 1 | Video stimulus used for eye-tracking analysis. This is an exemplary screenshot of the classroom and areas of interest (AOI) used. AOIs are only marked for the purposes of illustration in this paper, and were not visible to participants. Faces are also blurred for presentation in the publication to ensure the protection of data privacy; faces were visible to participants during the study and when drawing the AOIs. Students were marked with letters that did not refer to any underlying profile: B, E, K, P, and T. This figure was previously published as "Video stimulus for eye movement analysis" by Seidel et al. (2020) and is licensed under CC BY 4.0. The original figure was changed by adding the Table in the lower part.

of the five target students. Moreover, the points for each profile were added up across all profiles. The participants could therefore receive between 0 and 5 points overall. Since each profile could only be assigned once, when four correct assignments were made, the fifth profile would result from exclusion, and the overall judgment accuracy score was recoded to range from 0 (no correct assignment) to 4 (only correct assignments).

Student Observation

To assess teachers' observation behavior we used eye movement data. To ensure high quality of these data, we set two thresholds. First, the tracking ratio had to be at least 90%. Second, the deviations on the horizontal x-axis and vertical y-axis during validation of the calibration process were not allowed to exceed 1° (Holmqvist et al., 2011). Due to these quality criteria, eye movement data were processed for $n = 32$ (74.4%) of the participants with an average tracking ratio of 96% and average deviations on the x-axis = 0.49° and y-axis = 0.56°. We defined each of the five target students as one dynamic area of interest (AOI) using the BeGaze software version 3.7 (SMI, 2017a) to capture eye movement related to each (**Figure 1**). The exactness of the drawn AOIs was ensured throughout the whole video by

making manual adjustments to the AOIs whenever needed, for example when a student leaned over toward their neighbor. To identify fixations in participants' eye movements, we used the default velocity-based algorithm as recommended by Holmqvist et al. (2011) and implemented in previous eye tracking studies (Wolff et al., 2016). Thus, the fixation count (i.e., number of fixations on one AOI) and the average fixation duration (i.e., the average length of one fixation within one AOI) were assessed in relation to each of the target students.

Student Cue Utilization

To assess the student cues utilized, we coded participants' open answers to the question of which cues they had observed and utilized to assess student characteristics profiles. This question was asked separately for each target student. Thus, participants could provide between zero and five answers because answering this question for every target student was voluntary. Nevertheless, the majority of participants ($n = 27$, 62.8%) had indicated cues for at least one student profile, resulting in 124 answers. These answers were equally distributed across the five profiles. Most answers were provided as lists of single student cues separated by semicolons or bullet points. Therefore, we chose these single student cues as units of analysis for coding, resulting in 376 units. Two researchers coded these units inductively, resulting in a final coding scheme of 26 single codes (**Table 1**) for which they reached a high interrater agreement ($\kappa = 0.93$ for segment comparison with a minimum code intersection rate of 95% at the segment level). Based on research on student engagement (Fredricks et al., 2004, 2016a,b; Appleton et al., 2008; Rimm-Kaufman et al., 2015; Sinatra et al., 2015; Lam et al., 2016; Chi et al., 2018; Böheim et al., 2020), we then clustered the single codes in five categories, namely (1) behavioral, (2) cognitive, and (3) emotional engagement, pointing toward the intensity of student engagement; (4) knowledge, which represents the content of student engagement; and (5) student confidence.

Data Analysis

Judgment Accuracy

To investigate student teachers' judgment accuracy, we applied a mixture of descriptive and non-parametric testing, first by visually inspecting the distribution of judgment accuracy scores. Second, a Friedman test for repeated measures with Dunn-Bonferroni post-hoc comparisons corrected for multiple comparisons was calculated to examine whether participants differed in their judgment accuracy among student profiles. This non-parametric procedure was chosen because accuracy scores on profile level could only take on the values of zero (incorrect assessment) and one (correct assessment). Thus, they deviated strongly from a normal distribution. Finally, we descriptively investigated which profiles student teachers interchanged most frequently with one another to determine which profiles were difficult to distinguish.

Observation of Students

To investigate whether student teachers with low and high judgment accuracy differed in fixation count and average fixation duration, we summed up the fixation counts and averaged the fixation durations for each target student. To investigate the effects of high vs. low judgment accuracy on these variables, we split the whole sample along the median of the overall accuracy score and identified the two subgroups accordingly (Iacobucci et al., 2015a,b). This resulted in a group of $n = 21$ with low judgment accuracy ($M_{\text{judgment accuracy}} = 1.14$, $SD = 0.73$) and a group of $n = 22$ with high judgment accuracy ($M_{\text{judgment accuracy}} = 3.27$, $SD = 0.46$). These two subgroups differed significantly from one another in their judgment accuracy [$T_{(33.35)} = -11.45$, $p < 0.001$]. High-quality eye tracking data were available for $n = 17$ low accuracy and $n = 15$ high accuracy student teachers. We calculated a series of unpaired t-tests to compare the number of fixations and the average fixation duration between student teachers with low and high judgment accuracy for each student profile (see **Table 2** for intercorrelations). Here, we used Bonferroni adjusted p-values for multiple testing to consider alpha error accumulation.

Student Cue Utilization

To identify which student cues participants used to assess student profiles, we inspected the relative frequencies of the inductively derived codes.

To investigate differences in cue utilization among student teachers with either high or low judgment accuracy, we compared whether the two groups used different combinations of student cues to infer student profiles. To achieve this, we again compared the median split subgroups. Of these, $n = 11$ student teachers with low judgment accuracy and $n = 16$ student teachers with a high judgment accuracy had provided at least one answer to the question of which student cues they had used to assess a specific student profile. To investigate how both groups of student teachers used cue combinations in this context, we applied an epistemic network analysis (ENA, Shaffer et al., 2009, 2016; Shaffer and Ruis, 2017; Csanadi et al., 2018; Wooldridge et al., 2018) using ENA Web Tool version 1.7.0 (Marquart et al., 2018). ENA is a graph-based analysis that allows for modeling of the structure and strength of co-occurrences of a relatively small number of elements in a network (see for examples of this kind of analysis Andrist et al., 2015; Csanadi et al., 2018; Wooldridge et al., 2018). In general, the size of network nodes corresponds to the importance of that element in the network, and the strength of the connection between two elements represents the frequency of their combination (see **Figure 2** for prototypical networks). Thus, ENA allowed us first, to investigate which single student cues co-occurred within answers to the question of which cues had been utilized, and second, whether teachers with low and high judgment accuracy differed in the way they combined student cues to assess student characteristics profiles.

ENA is based on three core entities. The first entity consists of the *codes* for which one wants to investigate co-occurrences. In this study, predominantly reported student cues were used as codes. Therefore, we included cues that accounted for more than 3% of all codings (see **Table 1** for the included cues). The second entity refers to the *unit of analysis*, which defines the object of ENA, for which we used judgment accuracy (low vs. high) and student profile (strong, struggling, overestimating, underestimating, and uninterested). Hence, within each group

TABLE 1 | Coding scheme for student cues and their relative frequency.

Category	Codes	Example cue	Relative frequency (in %)
INTENSITY OF ENGAGEMENT			
Behavioral	Participation	Writes a lot	**5.9**
	No participation	Does not work on the task	**5.0**
	Frequent hand-raisings	Raises hand very frequently	**9.4**
	No/few hand-raisings	Seldomly raises hand	**10.6**
	Fast working	Begins reading quickly	2.6
	Slow working	Last to turn over worksheet	2.6
	Following gaze	Looked at teacher	0.6
	Digressive gaze	Frequently allows gaze to wander	**5.0**
	Interacts with classmates	Communicates with neighbor	0.6
	Does not interact with classmates	Has not participated in discussion with neighbor	0.6
	Inconspicuous	Very quiet	2.4
	Otherwise involved	Keeps playing with her hair	**7.9**
Cognitive	Attentive	Has always followed the lecture	2.9
	Inattentive	Is rarely attentive in class	**3.8**
	Concentrated	Work is predominantly concentrated	1.5
Emotional	Interested	Shows interest	1.5
	Uninterested	Seems uninterested	2.1
	Bored	Appears to be bored	1.5
CONTENT OF ENGAGEMENT			
Knowledge	High quality of verbal contributions	Makes good contributions	**10.9**
	Low quality of verbal contributions	Provides weak responses	**5.6**
	Understanding of topic	Seems to comprehend	**3.2**
	Problems with understanding	Has to erase frequently	2.4
	Helps classmates	Provides help to classmates	2.6
	Receives help	Gets help from neighbor	1.2
CONFIDENCE			
	Confident	Confidence in providing answers	2.1
	Unconfident	Nervous laughter	**5.6**

Predominant cues, which were included in the epistemic network analysis, are marked in bold. Example cues are translated from German to English.

TABLE 2 | Intercorrelations for fixation count and average fixation duration across student profiles.

	1	2	3	4	5	6	7	8	9
FIXATION COUNT									
1. Strong									
2. Struggling	0.44**								
3. Overestimating	0.32**	0.29							
4. Underestimating	0.44**	0.30	0.54**						
5. Uninterested	0.54**	0.34*	0.48**	0.40**					
AVERAGE FIXATION DURATION									
6. Strong	0.25	0.17	0.00	0.06	0.28				
7. Struggling	0.17	0.22	−0.10	−0.11	0.21	0.82**			
8. Overestimating	0.09	0.11	−0.05	−0.09	0.06	0.70**	0.77**		
9. Underestimating	0.09	0.15	0.05	0.05	0.24	0.79**	0.77**	0.73**	
10. Uninterested	0.21	0.11	−0.02	−0.07	0.22	0.80**	0.78**	0.77**	0.82**

*$p < 0.05$, ** $p < 0.01$. Average fixation duration in milliseconds.*

of low and high judgment accuracy, one network per student profile was constructed. Third, *stanza* determines the proximity that codes must have to one another in order to be considered as co-occurring. In our case, each of the 124 answers to the question of which cues student teachers had utilized to assess the profile of student B/E/K/P/T were defined as stanzas. Hence, ENA took

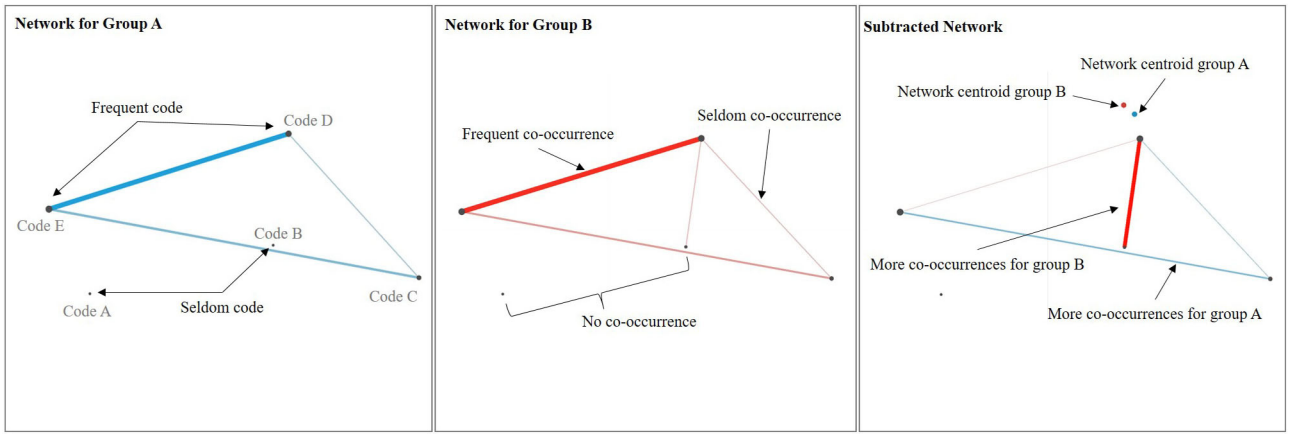


FIGURE 2 | Prototypical networks from epistemic network analysis. These are prototypical networks for two groups, A (blue) and B (red). Each node of the network represents a code. The size of the nodes represents the occurrence of single codes. The bigger a node, the more often the respective code appeared. The thickness of the connections between the nodes represents the co-occurrences of codes. The thicker the connecting line, the more often the respective codes co-occurred. If codes did not co-occur, they remain unconnected. A subtracted network illustrates the comparison between networks A and B. The color and strength of the connecting lines indicate for which group specific codes co-occurred more often. A network centroid summarizes the structure of a network in one point.

only code co-occurrence within single answers into account. For example, to create a network for the strong profile within the group with a high judgment accuracy, the only cues used were those coded for that particular group's responses when asked which cues they had used to assign a profile to student "P."

To create networks, one adjacency matrix was created for each stanza, indicating whether each of all possible code combinations was present or absent in a particular stanza. In our case, for every answer, one matrix was constructed to indicate whether combinations of student cues were present or absent. These adjacency matrixes were then accumulated for each unit, representing the number of stanzas for which each student cue combination was present. Each cumulative adjacency matrix was then converted into an adjacency vector. Thus, a high-dimensional space is created in which each dimension represents a specific combination of student cues. These vectors may vary in their length, because for some student profiles more or fewer answers were available than for others. To account for these differences, the adjacency vectors were spherically normalized to represent relative frequencies of student cue combinations. Next, the high-dimensional space was reduced to a low-dimensional projected space via mean rotation—to maximize the differences between the two groups of student teachers with low and high judgment accuracy—and singular value decomposition (SVD), a method similar to principal component analysis that reduces the number of dimensions to those that explain most variance in the data. ENA represents each network in the low-dimensional projected space both by a single point, which locates that unit's network centroid (a summary of the structure of its connections), and a weighted network graph. To visualize the network graphs, we chose mean rotation as the x-axis and the singular value that explains most of the variance in the data as the y-axis. The network graphs were then visualized using nodes and edges. Nodes correspond to the student cues. Their position is fixed due to an optimization routine that minimizes deviations between the plotted points and the respective network centroids. A correlation was estimated for the relation between the centroids and the projected points as a measure of model fit. Edges represent the relative frequency of co-occurrences of the cues (Andrist et al., 2015; Shaffer et al., 2016; Shaffer and Ruis, 2017).

These ENA characteristics allow quantitative and qualitative comparisons of networks for single participants or groups, as is the case in our study. Specifically, we compared the location of the network centroids for the two groups of low and high judgment accuracy student teachers with *t*-tests along the x- and y- axes and inspected so-called subtracted networks (see **Figure 2** for a prototypical subtracted network). Subtracted networks visualize the differences between two networks and compare which code co-occurrences were more frequent in which network. The edges are color-coded accordingly and indicate in which group the code combination occurred more frequently. This visualization allows a qualitative comparison of code co-occurences between our two groups but does not test whether these differences are significant.

RESULTS

Judgment Accuracy

To investigate student teachers' judgment accuracy the distribution of student teachers' overall judgment accuracy score is depicted in **Figure 3**. Student teachers had an average judgment accuracy score of 2.23 ($SD = 1.23$) across all profiles. This distribution indicates that student teachers differ substantially in their judgment accuracy. Roughly one half of the participants assessed three or five student profiles correctly and gained an accuracy score of three or four points, while the other half showed difficulties in accurately assigning student profiles.

A significant Friedman test [$\chi^2(4) = 25.53$, $p < 0.001$] indicated that student teachers' judgment accuracy differed among student profiles. As shown in **Figure 4**, the following descending order was identified for average judgment accuracy at profile level: Uninterested, struggling, underestimating, strong,

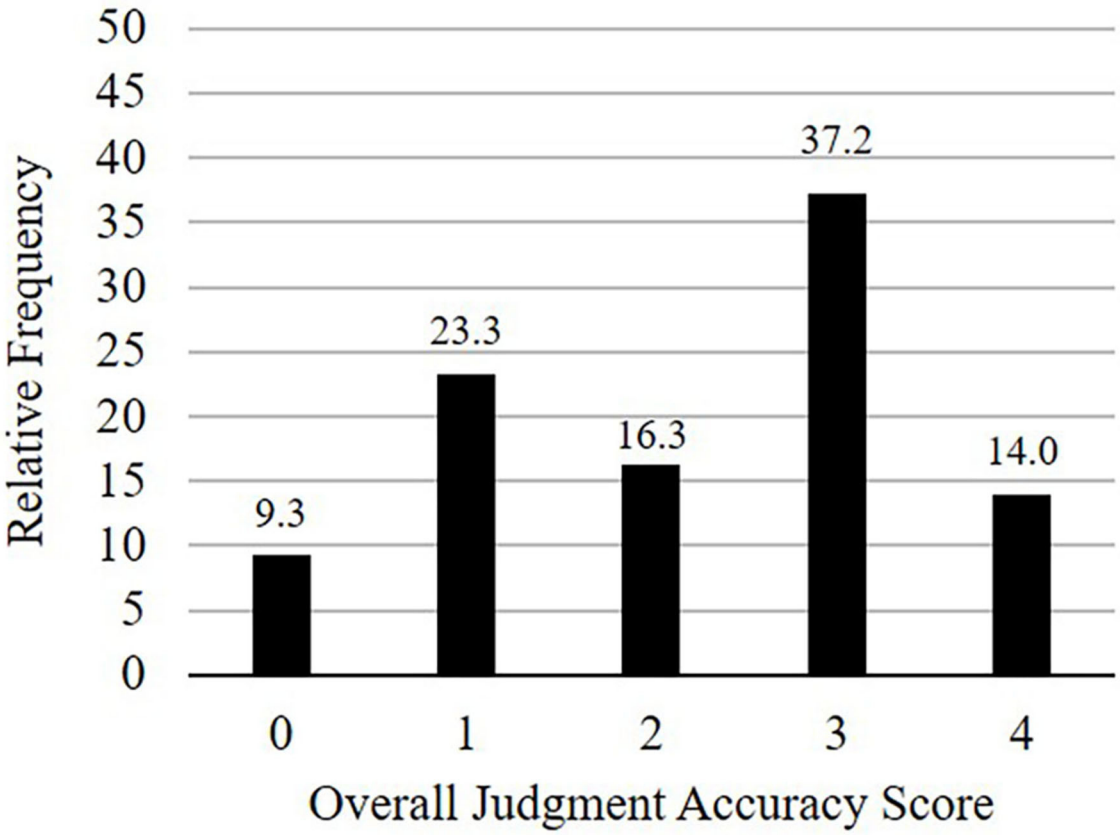


FIGURE 3 | Distribution of overall judgment accuracy score.

and overestimating. Results of the post-hoc tests (**Table 3**) suggest that student teachers tended to assign students to the uninterested ($M = 0.67$, $SD = 0.47$) and the struggling profile ($M = 0.65$, $SD = 0.48$) with a similarly high accuracy. Moreover, these profiles were clearly judged more accurately than the underestimating ($M = 0.40$, $SD = 0.49$), the strong ($M = 0.35$, $SD = 0.48$), and the overestimating ($M = 0.30$, $SD = 0.46$) ones. However, none of these differences reached significance when correcting for multiple comparisons.

Regarding student teachers' difficulties in distinguishing students with certain profiles from one another, our descriptive analysis yielded three main findings (**Figure 5**). First, the strong and overestimating profiles were most often "mixed up," or interchanged. The strong profile was even more often assigned to the overestimating student (65.1%) than to the strong one (34.9%). Similarly, the overestimating profile was more often assigned to the strong student (51.2%) than to the overestimating one (30.2%). Second, the struggling and the underestimating profiles were also frequently interchanged. Of the participants, 20.9% assigned the underestimating profile to the struggling student and 32.6% assigned the struggling profile to the underestimating student. Third, the profiles of lower motivational-affective characteristics—struggling, underestimating, and uninterested—were also sometimes confused. These findings suggest that motivational-affective characteristics seem to outshine cognitive characteristics; profiles with similar motivational-affective characteristics were frequently interchanged with one another (for example strong and overestimating) while profiles with different motivational-affective characteristics were rather clearly distinguished (for example strong and struggling).

Differences in Eye Movements

Descriptive statistics for student teachers' fixation counts and average fixation durations, as well as *t*-test results regarding differences in these variables between high and low judgment accuracy, are presented in **Table 4**. When comparing means for both groups descriptively, student teachers with a high judgment accuracy displayed the anticipated pattern. They had higher fixation counts on each student besides the struggling one, and showed shorter average fixation durations on each of the target students than student teachers with low judgment accuracy. *T*-tests indicated that student teachers with high judgment accuracy had more fixations on the overestimating [$t_{(30)} = -2.72$, $p = 0.011$] student and showed shorter average fixation durations for the strong [$t_{(30)} = 2.21$, $p = 0.035$] and underestimating student [$t_{(30)} = 2.52$, $p = 0.017$]. However, when adjusting for multiple comparisons, these differences were no longer significant. According to this, there are only minimal differences in the expected direction between the two groups that exist on a

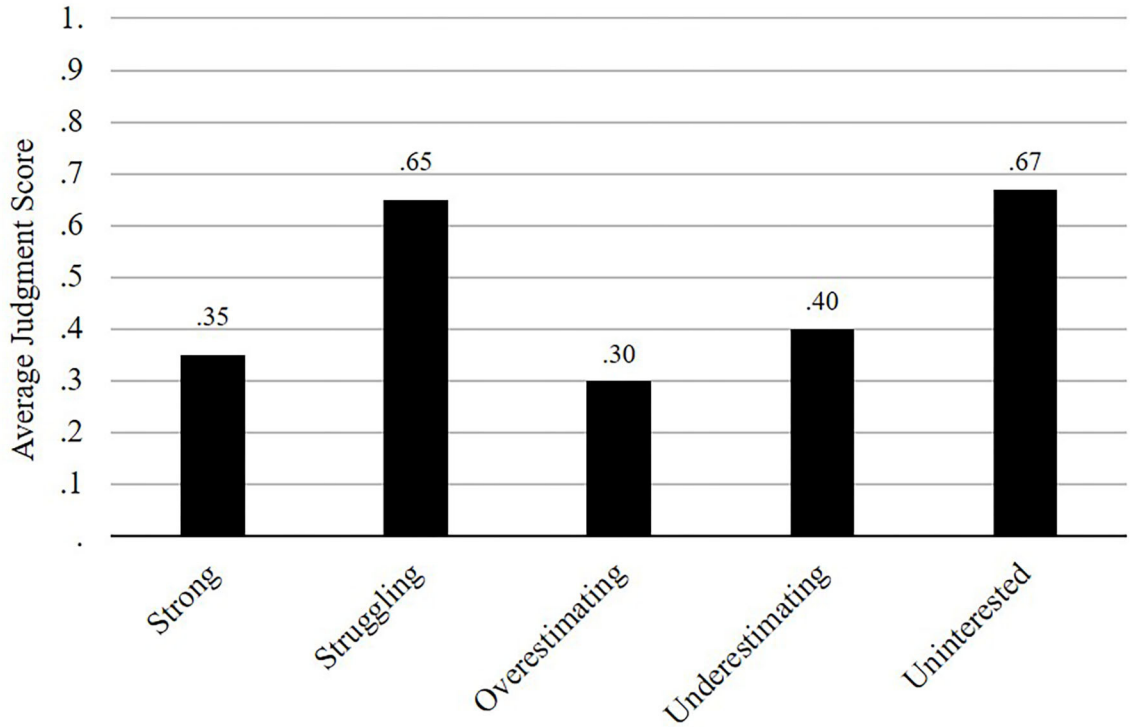


FIGURE 4 | Average judgment accuracy across student profiles.

TABLE 3 | *Post-hoc* tests comparing judgment accuracy among student profiles.

		1. Strong				2. Struggling				3. Overestimating				4. Underestimating			
	M (*SD*)	*z*	*p*	*adj. p*	*r*	*z*	*p*	*adj. p*	*r*	*z*	*p*	*adj. p*	*r*	*z*	*p*	*adj. p*	*r*
1. Strong	0.35 (0.48)																
2. Struggling	0.65 (0.48)	−0.756	0.027	0.267	0.12												
3. Overestimating	0.30 (0.46)	0.116	0.733	1.000	0.02	0.872	0.011	0.105	0.13								
4. Underestimating	0.40 (0.49)	−0.116	0.733	1.000	0.02	0.640	0.061	0.607	0.10	−0.233	0.495	1.000	0.03				
5. Uninterested	0.67 (0.47)	−0.814	0.017	0.170	0.12	−0.058	0.865	1.000	0.01	−0.930	0.006	0.064	0.14	−0.698	0.041	0.408	0.11

Adj. p, Bonferroni adjusted.

purely descriptive level and inferential statistics do not support our assumptions.

Differences in Student Cue Utilization

Student teachers reported a variety of cues, referring to the intensity of student behavioral, cognitive, and emotional engagement (18 cues in total), as well as to level of knowledge as the content of student engagement (6 cues in total), and to confidence (2 cues; **Table 1**). Thus, student teachers reported the use of more diverse cues concerning intensity of engagement than content of engagement. Relative frequencies demonstrated predominantly used cues were observations of general class participation, hand-raising behavior, preoccupation with things other than the lecture, and inattention, as well as the quality of verbal contributions, general understanding of the subject matter, and lack of confidence. Therefore, student teachers seemed to focus on diagnostic student cues when assigning student profiles.

For the epistemic network analysis, **Table 5** gives a descriptive overview on the usage of the different student cues between participants with a low and high judgment accuracy across the five student profiles.

Our epistemic network model had a good fit with the data with Spearman and Pearson correlation being equal to 1.00 both for our x-axis and y-axis. **Figure 6** presents a comparison of the networks between student teachers with high and low judgment accuracy. It depicts the network centroids for both groups as squares and for each profile as points. The closer the centroids are located to one another, the more similar the network structures. In our case, the centroids of student profiles that were often interchanged are located more closely to one another than those

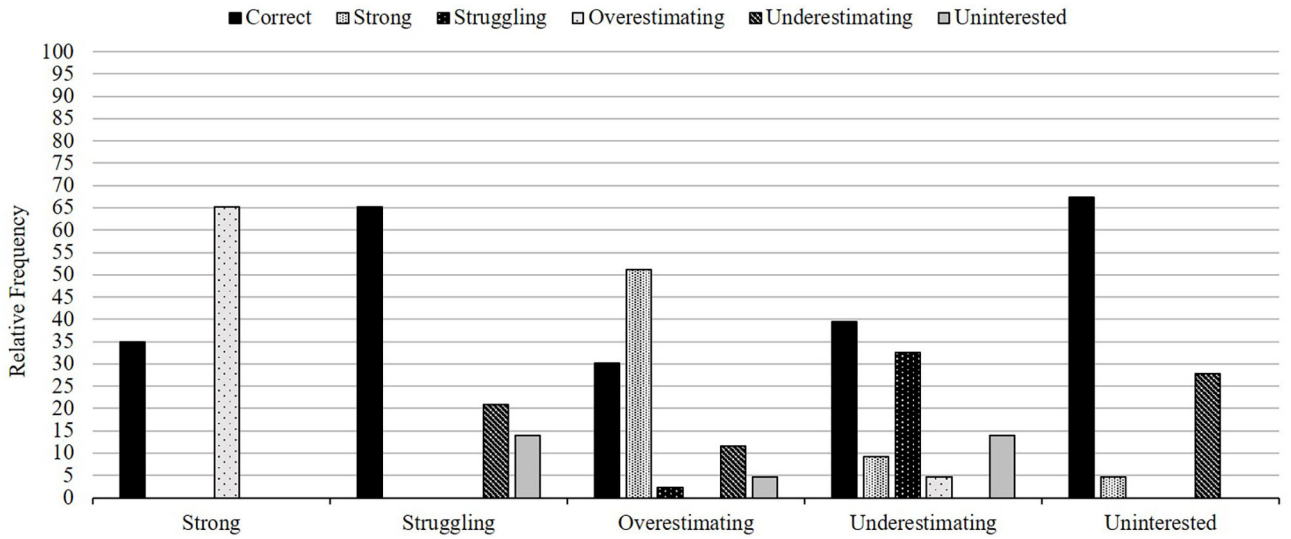


FIGURE 5 | Relative frequencies of correct assessments (black columns) and interchanged student profiles (gray shaded columns).

TABLE 4 | *T*-test comparisons for fixation count and average fixation duration for student teachers with high and low judgment accuracy.

	LA		HA		*T (df)*	*p*	95% CI	Adj. *p*	Cohen's *d*
	M	*(SD)*	*M*	*(SD)*					
FIXATION COUNT									
Strong	142.12	(33.67)	158.40	(30.49)	−1.43 (30)	0.164	[−39.59; 7.03]	1.00	−0.51
Struggling	147.76	(48.86)	141.33	(34.00)	0.43 (30)	0.673	[−24.37; 37.23]	1.00	0.15
Overestimating	121.53	(32.82)	150.27	(25.96)	−2.72 (30)	0.011	[−50.31; −7.16]	0.11	−0.97
Underestimating	177.06	(46.09)	222.40	(77.31)	−2.04 (30)	0.050	[−90.65; −0.03]	0.50	−0.71
Uninterested	157.00	(36.13)	162.00	(44.00)	−0.35 (30)	0.060	[−33.94; 23.94]	0.60	−0.12
AVERAGE FIXATION DURATION									
Strong	356.45	(83.20)	303.09	(45.24)	2.21 (30)	0.035	[4.04; 102.68]	0.35	0.80
Struggling	411.38	(110.68)	346.11	(85.37)	1.85 (30)	0.074	[−6.84; 137.38]	0.74	0.66
Overestimating	409.01	(88.77)	366.35	(124.97)	1.12 (30)	0.270	[−34.89; 120.21]	1.00	0.39
Underestimating	375.87	(71.95)	322.48	(41.64)	2.52 (30)	0.017	[10.16; 96.61]	0.17	0.91
Uninterested	407.52	(109.01)	365.80	(71.85)	1.26 (30)	0.218	[-25.94; 109.38]	1.00	0.45

LA, Low judgment accuracy; HA, High judgment accuracy; Adj. p, Bonferroni adjusted. Average fixation duration in milliseconds.

that were not frequently interchanged. Specifically the centroids of the struggling, underestimating, and uninterested profiles are in local proximity, as are those of the strong and overestimating profiles, indicating that interchanges between student profiles are reflected in similar network structures. For the comparison of network centroids of student teachers with low and high judgment accuracy, we found significant differences along the x-axis [$t_{(5.21)} = -3.60$, $p = 0.01$, $d = 2.28$]. Deviation on the y-axis was non-significant [$t_{(8.00)} = 0.00$, $p = 1.00$, $d = 0.00$]. Thus, in general, the networks for both groups differed from one another.

The qualitative inspection of subtracted networks for each student profile, however, provided a detailed picture of how both groups differed in their use of combinations of student cues. Networks from student teachers with low and high judgment accuracy for all student profiles are shown in **Figures 7–11**. Each of these figures presents networks for high judgment accuracy participants in part (a), networks for low judgment accuracy participants in part (b), and subtracted networks for group comparison in part (c). For the purposes of our analyses, we first described for each student profile the dominant pattern of student cues for participants with low and high judgment accuracy and make relations of the student cues to underlying motivational and cognitive characteristics. Second, we inspected the subtracted networks to identify differences in the utilization of student cues, meaning differences in which student cues were reported in combination with one another, between student teachers with high and low judgment accuracy.

Networks for the strong student profile are shown in **Figure 7**. Both, student teachers with high and low judgment accuracy, focused heavily on a combination of two student cues to diagnose

TABLE 5 | Absolut and relative frequencies of utilized student cues for student teachers with high and low judgment accuracy across student profiles.

	Strong			Struggling			Overestimating			Underestimating			Uninterested		
	All	HA	LA	All	HA	LA	All	HA	LA	All	HA	LA	All	HA	LA
Number of answers	25	14	11	24	14	10	25	15	10	24	14	10	26	15	11
Student cue															
Intensity of engagement															
Participation	7 (28)	5 (38)	2 (18)	-	-	-	8 (32)	6 (50)	2 (20)	5 (21)	4 (29)	1 (10)	-	-	-
No participation	-	-	-	3 (13)	3 (21)	-	-	-	-	5 (21)	3 (21)	2 (20)	9 (35)	4 (27)	5 (45)
Frequent hand-raisings	16 (64)	9 (69)	7 (64)	1 (4)	-	1 (10)	14 (56)	9 (75)	5 (50)	1 (4)	1 (7)	-	-	-	-
No/few hand-raising	3 (12)	-	3 (27)	10 (41)	4 (29)	6 (60)	1 (4)	1 (8)	-	17 (71)	11 (79)	6 (60)	5 (19)	1 (7)	4 (36)
Digressive gaze	1 (4)	-	1 (9)	-	-	-	-	-	-	1 (4)	-	1 (10)	15 (58)	10 (67)	5 (45)
Otherwise involved	2 (8)	-	2 (18)	1 (4)	-	1 (10)	1 (4)	1 (8)	-	3 (13)	1 (7)	2 (20)	20 (77)	12 (80)	8 (73)
Inattentive	1 (4)	1 (8)	-	-	-	-	1 (4)	1 (8)	-	4 (17)	1 (7)	3 (30)	7 (27)	3 (20)	4 (36)
Content of Engagement															
High quality verbal contributions	11 (44)	8 (62)	3 (27)	4 (17)	-	4 (40)	16 (64)	8 (67)	8 (80)	3 (13)	3 (21)	-	3 (12)	-	3 (27)
Low quality verbal contributions	3 (12)	2 (15)	1 (9)	14 (58)	10 (71)	4 (40)	1 (4)	1 (8)	-	1 (4)	-	1 (10)	-	-	-
Understanding of topic	3 (12)	2 (15)	1 (9)	-	-	-	1 (4)	1 (8)	-	4 (17)	3 (21)	1 (10)	3 (12)	2 (13)	1 (9)
Confidence															
Unconfident	-	-	-	12 (50)	8 (57)	4 (40)	1 (4)	1 (8)	-	5 (21)	4 (29)	1 (10)	1 (4)	-	1 (9)

HA, high judgment accuracy; LA, low judgment accuracy; -, student cue was not used. Numbers in brackets show the relative frequencies, which is the percentage of answers containing the respective student cue. The indication of student cues was a voluntary part of the study. Due to this reason, the number of answers differs between the student profiles within groups of high and low judgment accuracy.

this profile—frequent hand raisings (intensity of engagement) and high quality of answers (content of engagement). As shown in the subtracted network [part (c) in **Figure 7**], the group of student teachers with a low judgment accuracy differed in that they also reported many other combinations of student cues of which some contradicted the strong student profile (e.g., no hand raisings and preoccupation with things other than the lecture). Hence, high accuracy student teachers seem to use predominantly combinations of student cues which are clearly pointing to a strong profile while low accuracy student teachers indicated many different cue combinations that did not clearly refer to the strong profile.

Regarding the struggling student profile (see **Figure 8**), both groups used a pattern of three student cues for their judgment—an unconfident appearance combined with the avoidance of hand-raisings (intensity of engagement) and low quality of verbal contributions (content of engagement). Differences in these patterns are shown in part (c) in **Figure 8**. Those with a high judgment accuracy took also into account that the student showed a low level of general participation in the learning activities while those with a low judgment accuracy seem to rate the quality of the verbal contributions as high using this as a cue for their judgment. Thus, student teachers with a high judgment accuracy utilized combinations that unambiguously indicate a struggling student profile whereas those with a low judgment accuracy may have made a false assessment by perceiving the answers as being of high quality which might not be indicative of a struggling profile.

For the overestimating student profile (see **Figure 9**), student teachers with high and with low judgment accuracy relied mostly on a combination of three student cues—frequent hand-raisings, general active class participation (intensity of engagement), and high-quality of answers (content of engagement). The combination of these student cues is not a diagnostic feature of an overestimating profile, but rather of a strong one. As shown in the subtracted network [part (c) in **Figure 9**], high accuracy student teachers also used a variety of other cue combinations. Some of them captured aspects of an overestimating profile, such as making low quality verbal contributions. These other combinations might be important for high accuracy student teachers to assess the overestimating profile correctly.

In terms of the underestimating student profile (see **Figure 10**), both groups utilized a pattern of four student cues: An unconfident appearance combined with avoidance of hand-raisings but high participation in learning activities (intensity of engagement), and understanding of the topic (content of engagement). The subtracted network [part (c) in **Figure 10**] illustrates that in comparison high accuracy student teachers focused more on the general high participation and high quality of contributions while those with a low judgment accuracy combined the student cue of no hand-raisings with many other cues. Again, the main pattern of high accuracy student teachers seems to be a diagnostic feature for the student profile to be diagnosed. The variety of cue combinations of the low accuracy group, on the other hand, did not allow for a conclusive profile assessment.

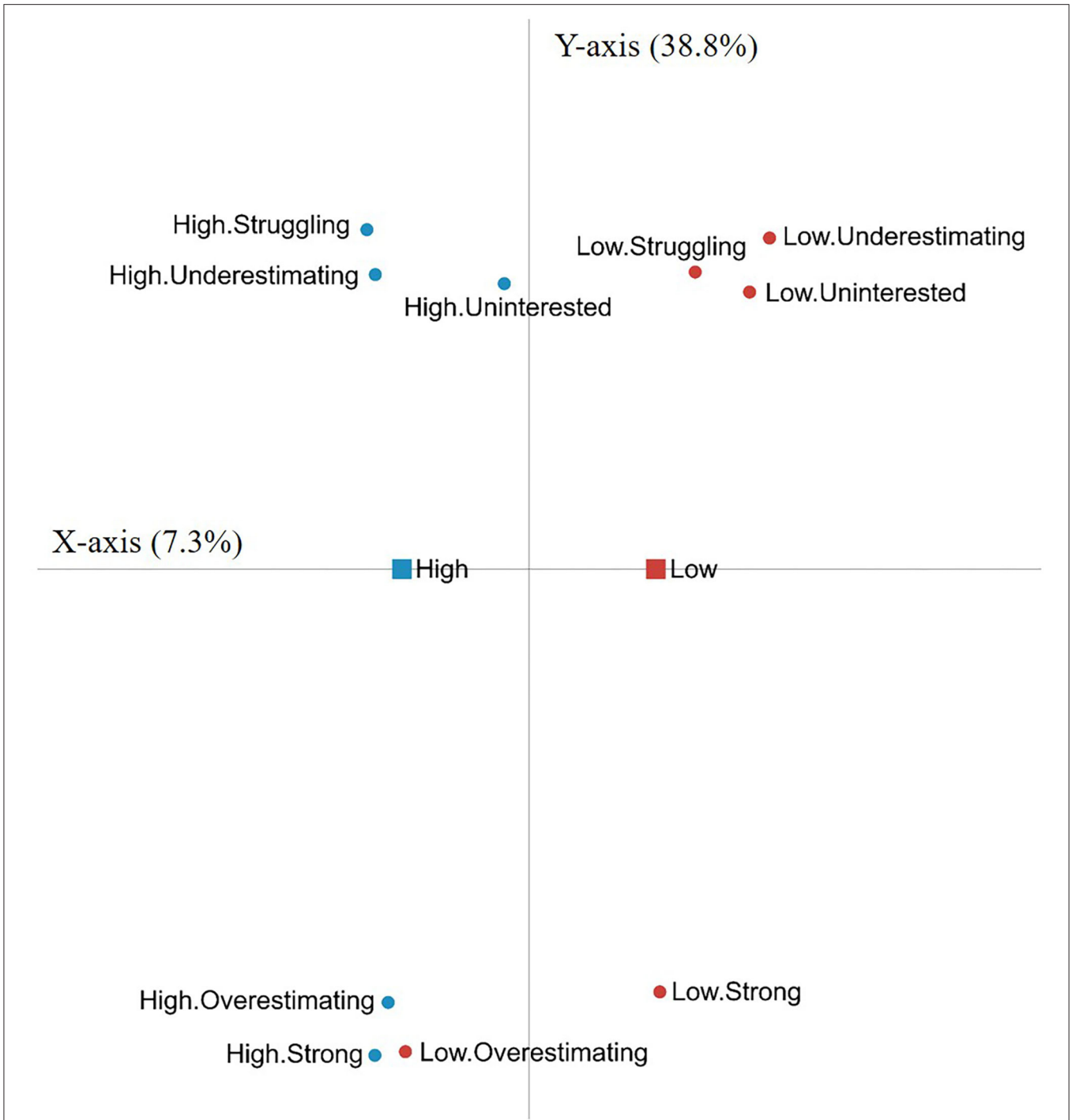


FIGURE 6 | Comparison of mean networks between student teachers with high and low judgment accuracy. High, High judgment accuracy (blue); Low, Low judgment accuracy (red). The figure presents the location of the centroids of the mean networks for HA and LA student teachers as squares. Centroids of the networks for each student profile are presented as points. The closer the centroids are located to one another, the more similar the network structures. The X-axis represents mean rotation, and the Y-axis represents the singular value that explains most of the variance in the data. Explained variance on the X-axis and Y-axis are 7.3 and 38.8%, respectively.

To identify the uninterested student profile (see **Figure 11**), both groups relied on a combination of two student cues: preoccupation with things other than the learning activities and an absent gaze (intensity of engagement). As shown in the subtracted network [part (c) **Figure 11**], the low judgment accuracy group also reported combinations of other cues

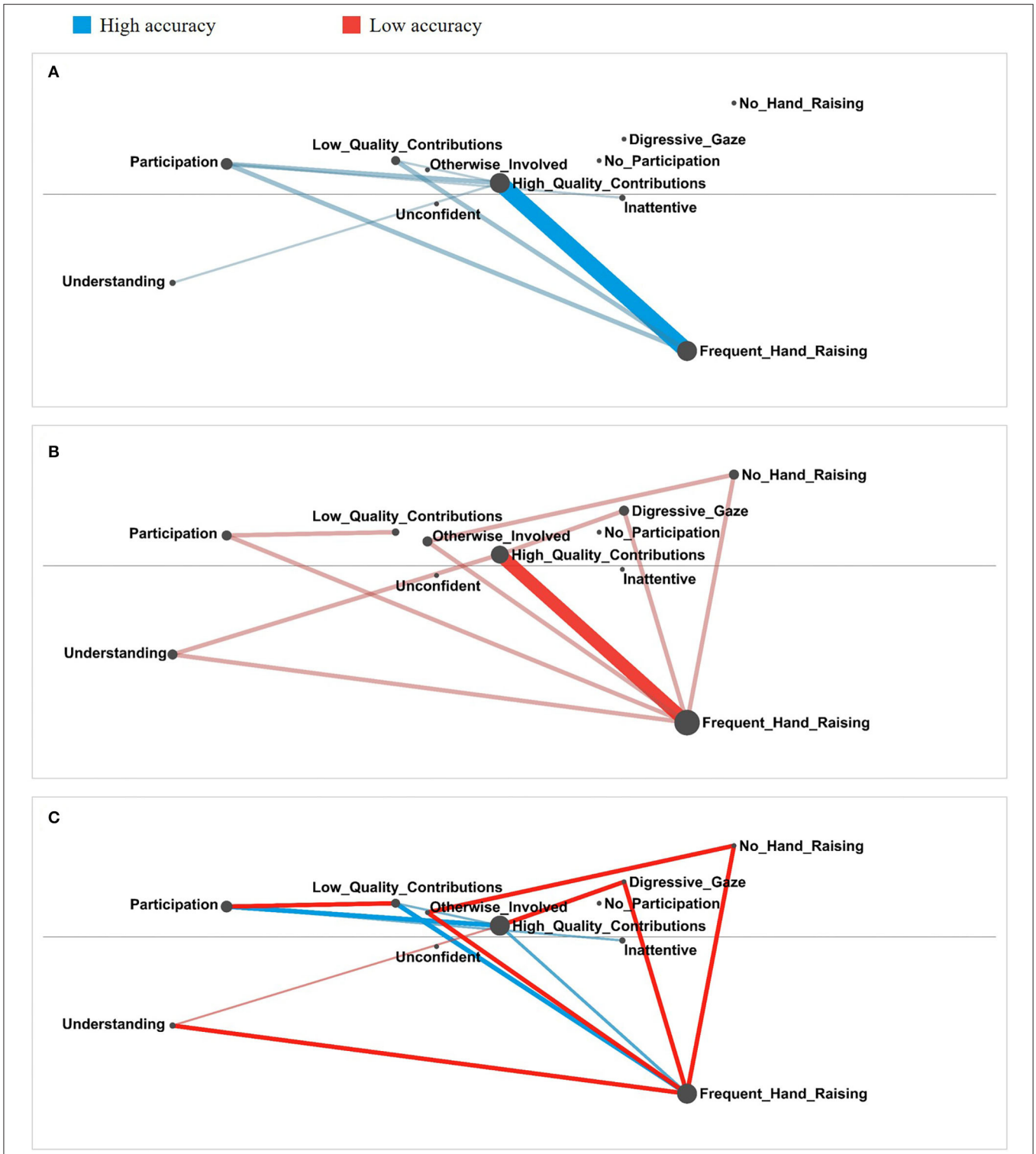


FIGURE 7 | Comparison of networks for strong student profile between student teachers with high and low judgment accuracy. **(A)** Depicts the network of the strong student profile for student teachers with high accuracy. **(B)** Depicts the network of the strong student profile for student teachers with low accuracy. **(C)** Shows the comparison of both networks. Here, the color of connections indicates which group utilized the respective code combination more frequently. In order to make the characteristics of the individual networks clearly visible and to make the differences visually easier to recognize, minimum edge weight and scale for edge weight were set to 0.1 and 1.3, respectively.

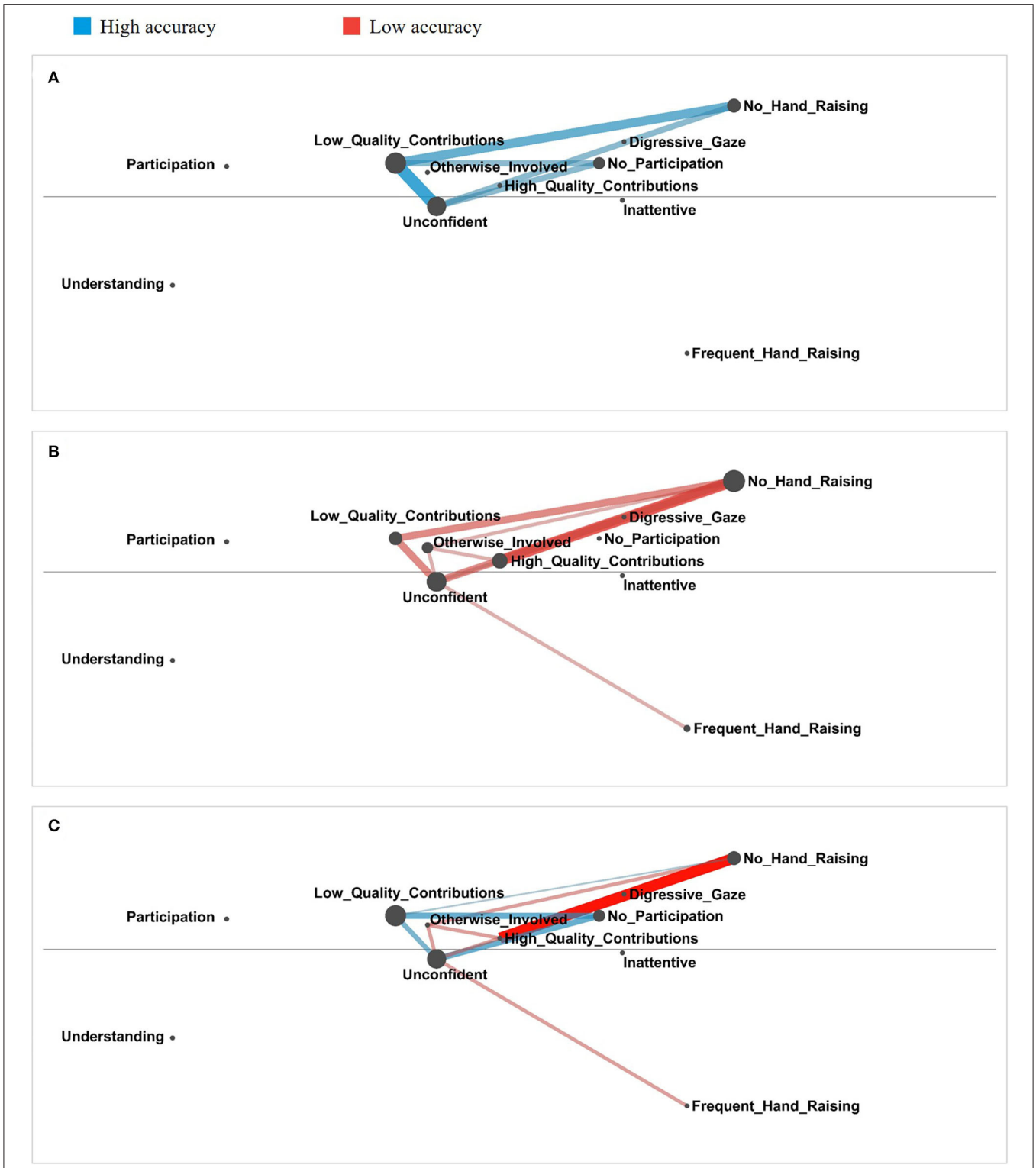


FIGURE 8 | Comparison of networks for struggling student profile between student teachers with high and low judgment accuracy. **(A)** Depicts the network of the struggling student profile for student teachers with high accuracy. **(B)** Depicts the network of the struggling student profile for student teachers with low accuracy. **(C)** Shows the comparison of both networks. Here, the color of connections indicates which group utilized the respective code combination more frequently. In order to make the characteristics of the individual networks clearly visible and to make the differences visually easier to recognize, minimum edge weight and scale for edge weight were set to 0.1 and 1.3, respectively.

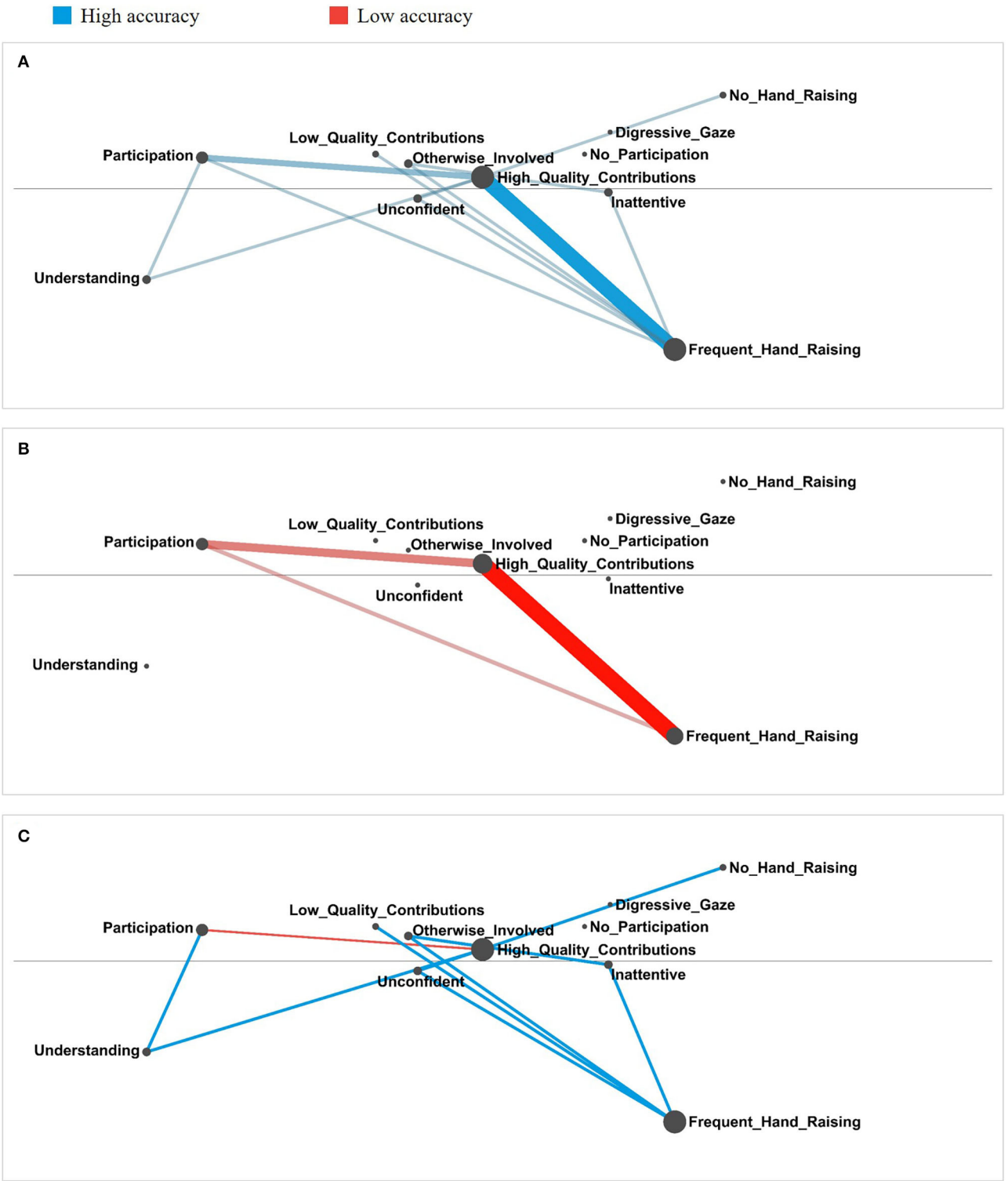


FIGURE 9 | Comparison of networks for overestimating student profile between student teachers with high and low judgment accuracy. **(A)** Depicts the network of the overestimating student profile for student teachers with high accuracy. **(B)** Depicts the network of the overestimating student profile for student teachers with low accuracy. **(C)** Shows the comparison of both networks. Here, the color of connections indicates which group utilized the respective code combination more frequently. In order to make the characteristics of the individual networks clearly visible and to make the differences visually easier to recognize, minimum edge weight, and scale for edge weight were set to 0.1 and 1.3, respectively.

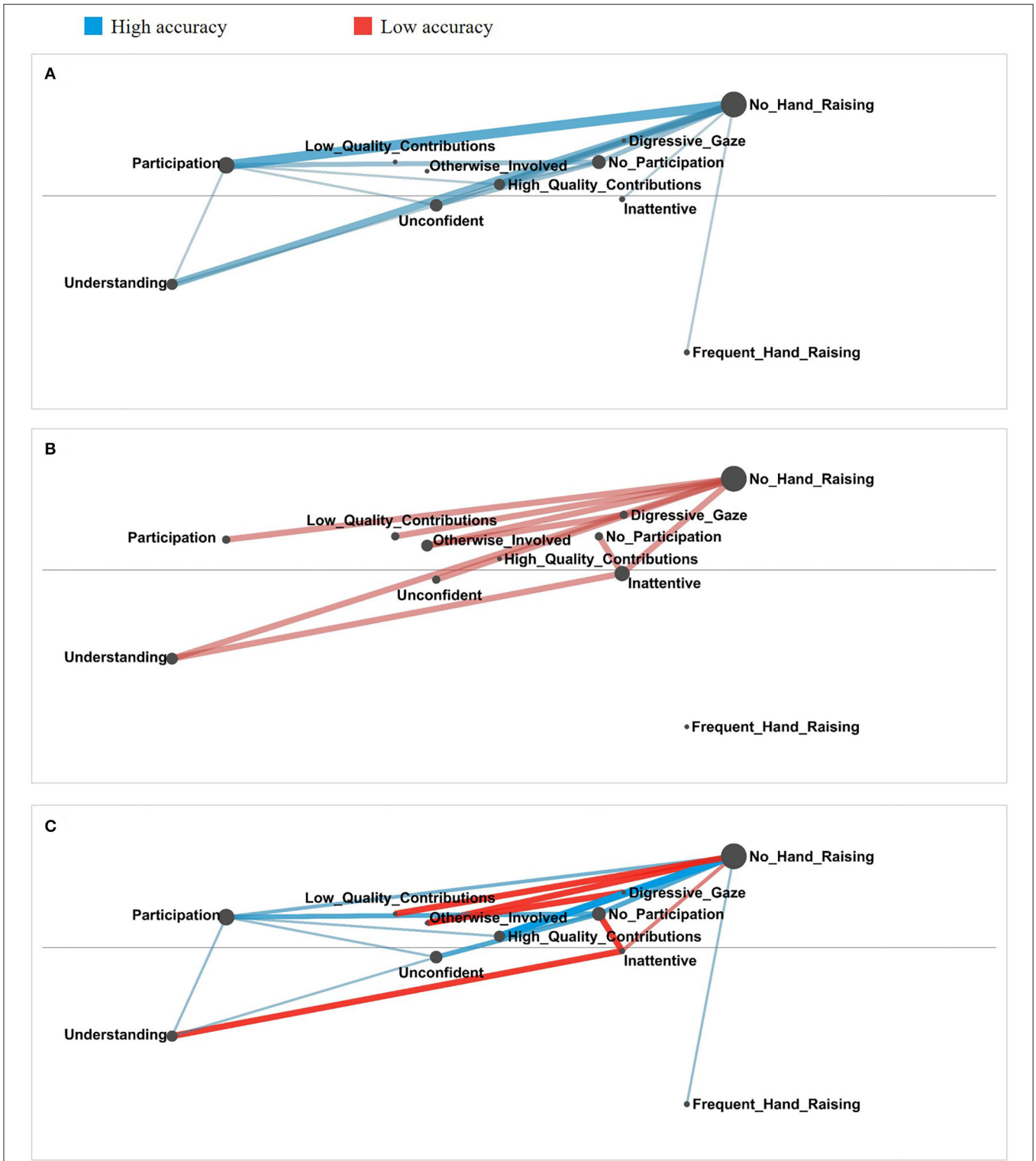


FIGURE 10 | Comparison of networks for underestimating student profile between student teachers with high and low judgment accuracy. **(A)** Depicts the network of the underestimating student profile for student teachers with high accuracy. **(B)** Depicts the network of the underestimating student profile for student teachers with low accuracy. **(C)** Shows the comparison of both networks. Here, the color of connections indicates which group utilized the respective code combination more frequently. In order to make the characteristics of the individual networks clearly visible and to make the differences visually easier to recognize, minimum edge weight and scale for edge weight were set to 0.1 and 1.3, respectively.

indicative for a low motivation (e.g., being inattentive). Hence, both groups utilized cue combinations that are diagnostic features of the uninterested profile. However, the high accuracy group focused on a combination specific to this profile, while the low accuracy group also included cues utilized to assess other profiles.

DISCUSSION

With the present study we aimed to connect teachers' judgment accuracy to preceding judgment processes. Therefore, we investigated student teachers' accuracy in assessing student profiles, which represent the diversity of students' cognitive and motivational-affective characteristics. Moreover, we explored the differences between student teachers with high and low judgment accuracy to shed light on the process of forming accurate judgments. Therefore, we considered eye movements when observing students as a behavioral activity associated with judgment processes and utilization of combinations of student cues as a cognitive activity.

Student Teachers' Differ in Judgment Accuracy of Student Characteristics

As part of our first research question, we investigated how accurately student teachers can judge student profiles overall, whether they vary in their accuracy across different student profiles, and which of the student profiles they interchange most frequently. In line with our assumptions and previous findings on student teachers' ability to interpret classroom events (Stürmer et al., 2016), the participating student teachers differed substantially in their judgment accuracy. One half assessed most of the student profiles correctly, while the other half struggled to do so. We had expected that student teachers would be more accurate in judging consistent profiles (strong, struggling) than inconsistent ones (overestimating, underestimating, uninterested) since teachers generally seemed to overestimate the level of consistency among their students (Huber and Seidel, 2018; Südkamp et al., 2018) and tended to intermingle cognitive and motivational-affective characteristics (Kaiser et al., 2013). Student teachers succeeded particularly well in recognizing the uninterested (inconsistent) and struggling (consistent) profiles, while they showed difficulties in identifying the strong (consistent), overestimating, and underestimating (inconsistent) student profiles. Hence, our assumption were partially disconfirmed.

This result can be explained with our exploration of the most frequent interchanges of student profiles. First, the uninterested and struggling student were less often interchanged with the other profiles resulting in a higher accuracy for these students. Regarding the uninterested profile, which was surprisingly accurately assessed despite its inconsistency, it can be assumed that this was because the student showed clear, easily observable cues that made judgment easy compared to the other profiles. Second, the strong and overestimating students were often interchanged, resulting in a low judgment accuracy for both profiles. Third, the student with the underestimating profile was often thought to be struggling or uninterested, also leading to low judgment accuracy. Hence, participants tended to interchange student profiles with similar levels of motivational-affective characteristics. This means that our student teachers were quite capable of assessing whether students were interested and feeling competent. However, assessing the level of cognitive characteristics appeared to be more challenging for them. This is somewhat surprising, as previous research findings have indicated that teachers have more difficulty in correctly assessing the motivation of their students than in assessing their achievement. So far it has been assumed that this is because student motivation is not directly observable, but must rather be inferred from the intensity of student engagement (Kaiser et al., 2013; Praetorius et al., 2017). For example, the level of student self-concept must be concluded based on hand-raising behavior (Böheim et al., 2020; Schnitzler et al., 2020). Moreover, it was previously assumed that classroom activities in general lack opportunities to observe student motivation (Kaiser et al., 2013). Thus, our results highlight student teachers' assessment competence. Although participants were unfamiliar with the students and had only limited opportunity to observe them via a relatively short video vignette as a proxy for real classroom teaching, some were already quite able to observe student engagement in a professional manner and thereby form correct judgments about student motivation. This is especially remarkable as our student teachers, only recently started to acquire declarative (pedagogical) knowledge (Shulman, 1986, 1987). Furthermore, they had only a few opportunities to gain the teaching experience necessary to the development of advanced knowledge structures that allow for the application of a professional knowledge base toward specific teaching situations (Carter et al., 1988; Berliner, 2001; Kersting et al., 2016; Lachner et al., 2016; Kim and Klassen, 2018).

However, the question remains as to why participants struggled to assess the level of cognitive characteristics. One possible reason may be that our video vignette did not provide enough opportunities to observe the content of student engagement. Another reason may be that the situations in which content of student engagement became salient were not sufficiently selective, and therefore differences in performance were not visible. Here, especially the observation of individual work that we did not show in the video could have been an important source of information containing other student cues about student cognitive characteristics. Hence, future research might systematically vary the available student cues with regard to the inclusion of information about intensity and content of engagement. Moreover, triangulation with qualitative analyses, like think-aloud protocols, could provide deeper insights into the conditions under which teachers experience either sufficient or limited availability of information about student cognitive and motivational-affective characteristics.

High Judgment Accuracy Relates Minimally to a More "Experienced" Pattern of Eye Movements When Observing Students

Our second research question investigated whether the eye movements of student teachers with a high judgment accuracy

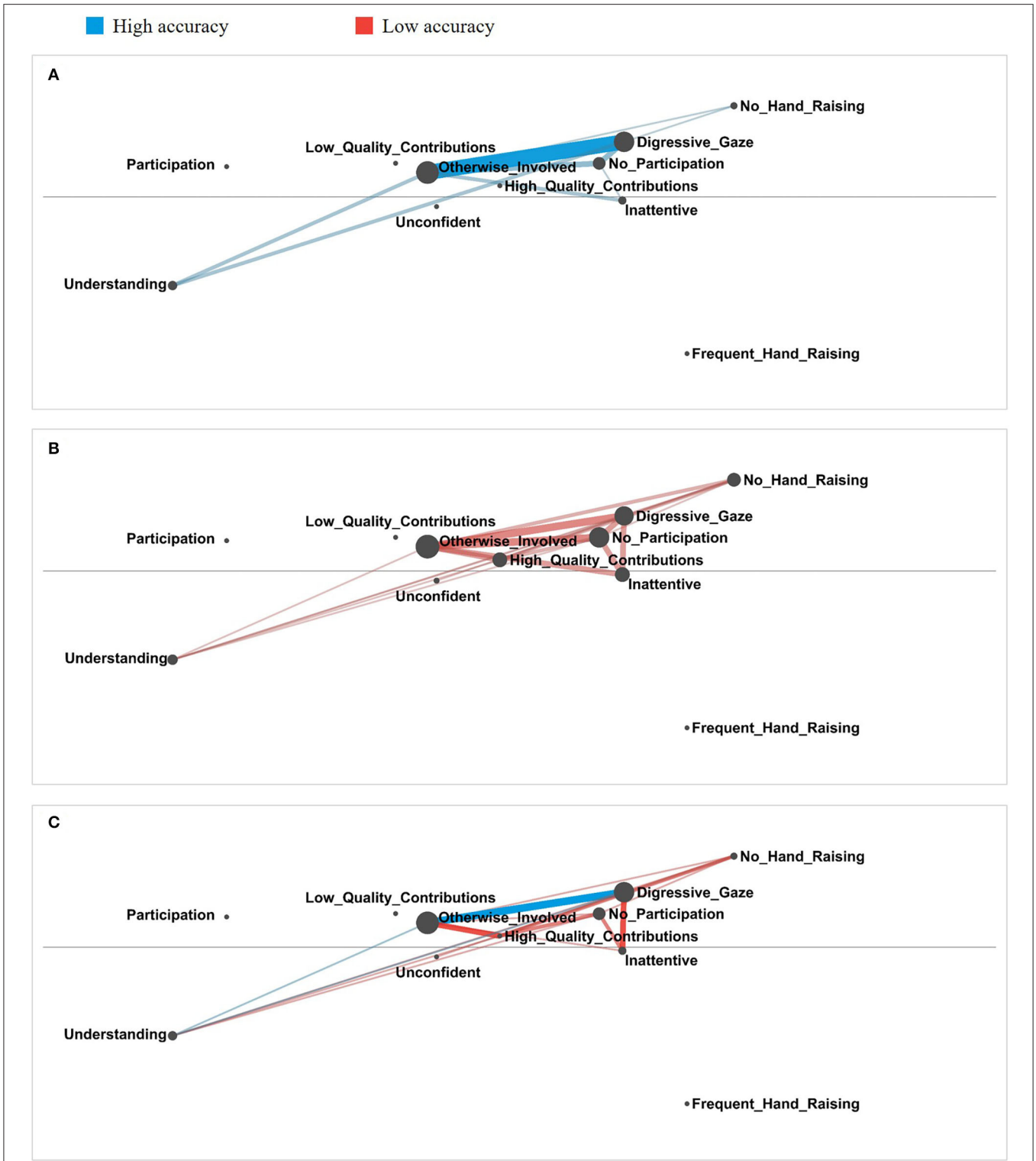


FIGURE 11 | Comparison of networks for uninterested student profile between student teachers with high and low judgment accuracy. **(A)** Depicts the network of the uninterested student profile for student teachers with high accuracy. **(B)** Depicts the network of the uninterested student profile for student teachers with low accuracy. **(C)** Shows the comparison of both networks. Here, the color of connections indicates which group utilized the respective code combination more frequently. In order to make the characteristics of the individual networks clearly visible and to make the differences visually easier to recognize, minimum edge weight and scale for edge weight were set to 0.1 and 1.3, respectively.

differed from those of student teachers with low judgment accuracy. To this end, we focused on the number of fixations and the average fixation duration. Here, a higher number of fixations and shorter average fixation duration represented an "experienced" pattern typical to expert teachers (Gegenfurtner et al., 2011; van den Bogert et al., 2014; Seidel et al., 2020). On a descriptive level we found the expected pattern, those student teachers with a low judgment accuracy showed slight tendencies to fixate the target students on average less often and for a longer time, a pattern typical for student teachers. In contrast, student teachers with a high judgment accuracy displayed an eye movement pattern minimally more similar to experienced teachers, with more fixations and a shorter average fixation duration. Overall, most of these differences were non-significant when correcting for multiple comparisons and therefore results must be interpreted with caution. Nevertheless, our findings make it likely that higher judgment accuracy might be associated with an "experienced" pattern of eye movements, which is an indicator for knowledge-driven observation and rapid information processing (Gegenfurtner et al., 2011; van den Bogert et al., 2014; Seidel et al., 2020). Our findings therefore show that eye movements are a relevant behavioral activity during judgment processes that allow for inferences about the accuracy of judgment formation. Thus, we expand upon the research of teachers' eye movements and emphasize its potential for investigation of issues other than those associated with classroom management. Since teaching is a vision-intense profession that requires teachers to regularly infer information from observing their classrooms (Carter et al., 1988; Gegenfurtner, 2020), the systematic investigation of eye movements might provide insights into different competencies of professional teachers. In terms of assessing teacher competence, future research might also investigate how (student) teachers distribute their gaze across different students and search for information. For example, do teachers with a higher judgment accuracy regularly check upon all of the students, or do they start to observe some students more intensively until they form a decision about their profile, and then move on to the next student for the purposes of profile assessment?

High Judgment Accuracy Relates to a Utilization of Particular Combinations of Diagnostic Student Engagement Cues

Our third research question was explorative in nature and followed the call of previous research to investigate which student cues teachers utilize to assess cognitive and motivational-affective characteristics, as well as their combination within individual students (Glock et al., 2013; Praetorius et al., 2017; Huber and Seidel, 2018; Brandmiller et al., 2020). Therefore, we investigated which student cues student teachers utilized to assess student characteristic profiles. Moreover, we aimed to identify differences in student cue utilization of student teachers with high and low judgment accuracy based on the assumptions of the lens model (Brunswik, 1956; Funder, 1995, 2012). In other words, accurate judgments of latent student characteristics depend on inference of intensity and content of student engagement as diagnostic cues for student motivation and cognitive characteristics. For the purposes of investigation, we applied the relatively new method of epistemic network analysis, which enabled us to gain detailed insights in how student teachers combined student cues to form judgments.

As outlined in our theory section (student) teachers need to observe and utilize student cues, which are diagnostic and provide information both about students' cognitive *and* motivational-affective characteristics, to assess student characteristic profiles accurately. According to our inductive coding of reported student cues, in general student teachers utilized diagnostic cues to assess student profiles. This means that they considered a mixture of student cues containing information about the intensity and content of student engagement, which relate back to student cognitive and motivational-affective characteristics. These were first and foremost the intensity and content of student behavioral engagement (Fredricks et al., 2004). Student teachers took into account in particular whether students showed general participation in learning activities, whether they raised their hands to contribute to classroom dialogue, and also considered the quality of students' verbal contributions frequently. That student teachers dominantly rely on such diagnostic student cues, contradicts previous research which showed that teachers also take into account misleading or unimportant information like student gender, ethnicity, immigration status, and SES in their assessment of student characteristics (Meissel et al., 2017; Praetorius et al., 2017; Garcia et al., 2019; Brandmiller et al., 2020). However, these studies used text vignettes to provide teachers with specific information about target students. Hence, our implementation of a video vignette as another proxy to everyday teaching, which contains rich information about students' engagement, might complement these prior findings, because teachers' utilization of student cues depends on availability of information, which differs between text vignettes and classroom videos (Funder, 1995, 2012). Thus, future research might systematically investigate the role of the stimulus (video or text vignette) and the amount and diversity of available information in teachers' use of diagnostic cues and ignorance of misleading ones. Additionally, our participants reported more diverse student cues with regard to the intensity of engagement than content of student engagement. This finding might explain why student teachers struggled to assess the level of student cognitive characteristics. The available student cues may have not contained diverse enough information to allow for a differentiation of student cognitive characteristics. For example, although student teachers considered the quality of student verbal contributions, they might not have provided deep insights into student knowledge because the video stemmed from an introductory lesson. The teacher's questions might have been rather easy so that most of the students were able to answer them correctly and could follow instructions. Other sources of information like students' solutions to mathematical tasks might contain more sufficient information to assess students' cognitive characteristics. Hence, upcoming studies might investigate which sources of information provide teachers with cues that allow for a differentiation between students in terms of cognitive and motivational-affective characteristics. Overall, here we provide

promising findings, in that student teachers are already able to observe and utilize diagnostic student cues.

Results from our epistemic network analysis pointed toward systematic differences in how student teachers with low and high judgment accuracy combined student cues. As we had expected, student teachers with a high judgment accuracy seemed to utilize combinations of student cues of intensity and content of engagement that were diagnostic features of particular student profiles. In contrast, those student teachers with a low judgment accuracy also relied on diagnostic student cues but seemed to utilize many different cue co-occurrences for each student profile, including misleading combinations. For example, to identify the struggling profile, both groups focused on combinations of an unconfident appearance, avoidance of hand-raisings (intensity of engagement), and low quality of answers (content of engagement). However, some student teachers with a low judgment accuracy seemed to rate the quality of the verbal contributions at the same time as high probably affecting their assessment and leading to incorrect judgments due to this misleading student cue. This overlaps with previous research, which reported that student teachers try to use as much information as possible to form judgments, while experienced teachers select the most relevant information (Böhmer et al., 2017). In this sense, student teachers with a high judgment accuracy seemed to have already developed a professional skill in that they were able to observe diagnostic student cues, utilize the relevant co-occurrences of these cues, and correctly infer student profiles. In terms of the development of teachers' professional vision, it can be assumed that student teachers with a high judgment accuracy are already able to apply their acquired declarative knowledge to assess student profiles from observation (Jacobs et al., 2010; Stürmer et al., 2013; Kersting et al., 2016; Lachner et al., 2016). Conversely, those with a low judgment accuracy struggle to recognize relevant information, as is quite typical of student teachers and beginning teachers (Carter et al., 1988; Berliner, 2001; Star and Strickland, 2008; Kim and Klassen, 2018; Keppens et al., 2019).

Another interesting finding resulted from our epistemic network analysis. Networks of student profiles that were frequently interchanged showed a rather similar structure. This means that difficulties in distinguishing the struggling, underestimating, and uninterested profiles from one another, as well as the strong and overestimating profiles, can be traced back to the utilization of similar combinations of student cues. This makes sense in part, since the interchanged profiles overlapped in their motivational-affective characteristics and therefore showed a similar intensity of engagement. However, differences in students' level of cognitive characteristics could have resulted in differentiated cue combinations of intensity of engagement and content of engagement, which would have allowed to distinguish the profiles. As outlined above, at this point it remains unclear whether this was due to the student cues contained in our video vignette or whether this is a general challenge for (student) teachers.

Our findings emphasize cognitive activities of judgment processes as a key to judgment accuracy. Teachers' utilization of student cues determines whether student teachers are successful in judging student characteristic profiles accurately. In this regard, epistemic network analysis seems to be a promising approach. Based on such an analysis, which visualizes the frequency of all co-occurrences of utilized student cues, it becomes evident that accurate judgments, difficulties with assessments, and interchanges of student profiles can be traced back to reliance on particular combinations of diagnostic student cues. Hence, epistemic network analysis allowed us to consider and investigate the complexity of everyday teacher judgment processes in which teachers are required to observe and interpret several pieces of information in combination to assess their students.

Practical Implications

Although our study has shown that some of the student teachers are already quite able to successfully assess student characteristics, a large number of them still struggles with this important task, which will later become a regular part of their professional everyday life. Furthermore, our results highlight the role of observing and using diagnostic student cues for accurate judgments. From this, several recommendations for teacher education can be derived. Teacher education should promote student teachers' declarative knowledge base with respect to learning relevant student characteristics, their intra-individual combination in consistent and inconsistent student profiles, and student cues that are diagnostic for these characteristics (that is intensity and content of engagement) as a foundation for the development of a professional vision in the context of assessing student characteristics (Stürmer et al., 2013). As the judgment of student characteristics requires student teachers to apply their knowledge toward teaching practice, they should receive support and instruction in how to do so in a step-by-step approach in which practice tasks are subsequently approximated to real-life teaching (Grossman et al., 2009). In this sense, the guided observation and reflection of classroom videos could be effective, like it has already been shown for the development of other areas of professional classroom perception (Star and Strickland, 2008; van Es and Sherin, 2008; Stürmer et al., 2013). Such practical experience might lead to changes in student teachers' perceptions of student cues, they could for example become more sensitive to the variance of participation among students leading to a more refined differentiation of whether students raise their hands often or seldom. Moreover, they might learn to identify, focus, and consistently use diagnostic student cues to assess the same student characteristic across several students (Nestler and Back, 2013). Besides this guided practice in teacher education courses, simulations have recently been discussed as a way to support teacher students in the development of their assessment skills (Chernikova et al., 2020a,b; see Codreanu et al., 2020 for an example concerning students' mathematical skills). In terms of observing students, it has been shown in other disciplines that the modeling of expert eye movements can help learners to develop effective eye movement patterns (Jarodzka et al., 2012, 2013). It is assumed that this can also be applied to teacher education although (intervention) studies are still pending (Gegenfurtner, 2020). Our findings are thus a good starting point for the development of appropriate teacher education programs.

Limitation

There are some methodological issues that need to be considered when interpreting the results. First, we used only one video as a stimulus. Thus, difficulties in distinguishing student profiles, the type of frequently used student cues, and differences in student cue utilization between student teachers may stem from specific features of the video. Hence, future studies might investigate whether our findings are replicable with other classroom videos and real-life teaching. Second, we did not systematically vary the available student cues, but rather used an every-day classroom video as a proxy for real classroom teaching so that student cue availability could be considered as "natural." In our case, we did not include a student individual work phase although this might have contained student cues that relate to student cognitive characteristics. This might have resulted in the rather weak performance of participants in judging student cognitive characteristics in comparison to other previous studies (Kaiser et al., 2013; Praetorius et al., 2017). Therefore, upcoming research might systematically investigate the influence of cue availability on judgment accuracy. Third, our video was relatively short, and student teachers were unfamiliar with the students. Although other studies showed that even very short videos are sufficient for accurate judgments (Praetorius et al., 2015), it might be interesting to systematically investigate whether familiarity with students influences judgment accuracy. Additionally, the comparisons of the eye movement patterns were based on median split of the whole sample, resulting in two relatively small subgroups. As a consequence, the tendencies in the expected direction toward differences in number of fixations and average fixation duration across student teachers with high and low judgment accuracy might have not reached significance. Therefore, differences might be underestimated within the present study and judgment accuracy might actually show a stronger connection to eye movements as a behavioral activity of judgment processes. Moreover, our participants provided the student cues on a voluntary base. This might have resulted in the presence of a bias and, above all, participants who are confident in their diagnostic accuracy might have worked on the task. Thus, it might have been likely that we included participants with a relatively higher judgment accuracy in our analysis on usage of student cues. However, we found no significant differences in the average diagnostic accuracy score across all student profiles between those participants who reported student cues ($M = 2.48$; $SD = 1.25$) and those who did not ($M = 1.81$; $SD = 1.1$); $t_{(41)} = -1.77$, $p = 0.085$. Additionally, the judgment accuracy of those participants included in our analysis on utilization of student cues spread across the whole range of accuracy scores, ranging between 0 and 4 points. Nevertheless, our results should be interpreted with this limitation in mind and differences in the patterns of student cue usage between student teachers with high and low judgment accuracy can unfortunately be less obvious in our study than they actually are. As a related issue, our analysis of student cue utilization is based on student teachers' self-reports. This might have resulted in a social desirability bias of their answers and would be one explanation why our findings contradict previous ones in which teachers relied on unimportant student cues such as gender or SES (Meissel et al., 2017; Praetorius et al., 2017; Garcia et al., 2019; Brandmiller et al., 2020). In this context, our results might not fully reflect potentially problematic cue usage of student teachers. Finally, we argue that student teachers with a high judgment accuracy successfully applied their acquired declarative knowledge to our specific situation, although they had only limited opportunities to connect their knowledge base to teaching experiences. However, we did not measure our participants' declarative knowledge. Thus, future research might elaborate on this issue and consider student cue utilization and eye movements as a mediator between teachers' declarative knowledge and judgment accuracy.

CONCLUSION

The present study was one of the first that aimed to connect judgment processes to judgment accuracy. Therefore, we considered student diversity, in the form of student characteristic profiles, as an assessment target; investigated eye movements as a behavioral activity; and looked at utilization of student cues as a cognitive activity; all in keeping with the lens model. The results advanced the understanding of teachers' accurate judgments. First, we identified a level of diversity among student teachers. A significant portion of the sample group was already quite successful in the complex task of assessing student profiles. The methodology of eye tracking indicates that this success tends to go along with a more "experienced" pattern of eye movements. The epistemic network analysis demonstrated the importance of using specific diagnostic student cues for high judgment accuracy. With this study, we have brought together research on the judgment process and on judgment accuracy. This allowed us to provide detailed insights into the processes of accurate judgments and is a necessity to understanding teaching as a vision-intense profession.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available from the corresponding author upon request.

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

KS organized the data collection and database, performed statistical analysis, and wrote the first draft of the manuscript. All authors contributed to the conception and design of the study, manuscript revision, and read and approved the submitted version.

FUNDING

The development of the ENA webtool was funded by the National Science Foundation (DRL-1661036, DRL-1713110), the Wisconsin Alumni Research Foundation, and the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison. The present research project was funded by the Deutsche Forschungsgemeinschaft (German Research Foundation, grant no. SE1397/7-3). The funders had no role in the study's design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

Andrist, S., Collier, W., Gleicher, M., Mutlu, B., and Shaffer, D. (2015). Look together: analyzing gaze coordination with epistemic network analysis. *Front. Psychol.* 6:1016. doi: 10.3389/fpsyg.2015.01016

APA American Psychological Association (2017). *Ethical Principles of Psychologists and Code of Conduct*. Washington, DC: American Psychological Association (APA). Available online at: https://www.apa.org/ethics/code

Appleton, J. J., Christenson, S. L., and Furlong, M. J. (2008). Student engagement with school: critical conceptual and methodological issues of the construct. *Psychol. Schools* 45, 369–386. doi: 10.1002/pits.20303

Back, M. D., and Nestler, S. (2016). "Accuracy of judging personality," in *The Social Psychology of Perceiving Others Accurately*, eds J. A. Hall, M. S. Mast, and T. V. West (Cambridge: Cambridge University Press), 98–124. doi: 10.1017/CBO9781316181959.005

Baumert, J., and Kunter, M. (2006). Stichwort: professionelle Kompetenz von Lehrkräften [Keyword: Professional competence of teachers]. *Zeitschrift für Erziehungswissenschaft* 9, 469–520. doi: 10.1007/s11618-006-0165-2

Baumert, J., and Kunter, M. (2013). "The COACTIV model of teachers' professional competence," in *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers: Results From the COACTIV Project,* eds M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, and M. Neubrand (New York, NY: Springer), 25–48. doi: 10.1007/978-1-4614-5149-5_2

Berliner, D. C. (2001). Learning about and learning from expert teachers. *Int. J. Educat. Res.* 35, 463–482. doi: 10.1016/S0883-0355(02)00004-6

Binder, K., Krauss, S., Hilbert, S., Brunner, M., Anders, Y., and Kunter, M. (2018). "Diagnostic skills of mathematics teachers in the COACTIV study," in *Diagnostic Competence of Mathematics Teachers: Unpacking a Complex Construct in Teacher Education and Teacher Practice*, eds T. Leuders, K. Philipp, and J. Leuders (Cham: Springer), 33–53.

Blömeke, S., Gustafsson, J.-E., and Shavelson, R. J. (2015). Beyond dichotomies. *Zeitschrift für Psychologie* 223, 3–13. doi: 10.1027/2151-2604/a000194

Böheim, R., Knogler, M., Kosel, C., and Seidel, T. (2020). Exploring student hand-raising across two school subjects using mixed methods: an investigation of an everyday classroom behavior from a motivational perspective. *Learn. Instruct.* 65:101250. doi: 10.1016/j.learninstruc.2019.101250

Böhmer, I., Gräsel, C., Krolak-Schwerdt, S., Höstermann, T., and Glock, S. (2017). "Teachers' school tracking decisions," in *Competence Assessment in Education: Research, Models and Instruments*, eds D. Leutner, J. Fleischer, J. Grünkorn, and E. Klieme (Cham: Springer International Publishing), 131–148. doi: 10.1007/978-3-319-50030-0_9

Böhmer, I., Hörstermann, T., Gräsel, C., Krolak-Schwerdt, S., and Glock, S. (2015). Eine Analyse der Informationssuche bei der Erstellung der Übergangsempfehlung: Welcher Urteilsregel folgen Lehrkräfte? [An analysis of the search for information when preparing the school tracking recommendation: What judgment rule do teachers follow?]. *J. Educat. Res. Online* 7, 59–81.

Boshuizen, H. P. A., Gruber, H., and Strasser, J. (2020). Knowledge restructuring through case processing: the key to generalise expertise development theory across domains? *Educat. Res. Rev.* 29:100310. doi: 10.1016/j.edurev.2020.100310

Brandmiller, C., Dumont, H., and Becker, M. (2020). Teacher perceptions of learning motivation and classroom behavior: the role of student characteristics. *Contemporary Educat. Psychol.* 51, 336–355. doi: 10.1016/j.cedpsych.2020.101893

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments. 2nd Edn.* Berkeley: University of California Press.

Carter, K., Cushing, K., Sabers, D., Stein, P., and Berliner, D. C. (1988). Expert-novice differences in perceiving and processing visual classroom information. *J. Teacher Educat.* 39, 25–31. doi: 10.1177/002248718803900306

Chernikova, O., Heitzmann, N., Fink, M. C., Timothy, V., Seidel, T., and Fischer, F. (2020a). Facilitating diagnostic competence in higher education – a meta-analysis in medical and teacher education. *Educat. Psychol. Rev.* 32, 157–196. doi: 10.1007/s10648-019-09492-2

Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., and Fischer, F. (2020b). Simulation-based learning in higher education: a meta-analysis. *Rev. Educat. Res.* 4, 499–541. doi: 10.3102/0034654320933544

Chi, M. T. H., Adams, J., Bogusch, E. B., Bruchok, C., Kang, S., Lancaster, M., et al. (2018). Translating the ICAP theory of cognitive engagement into practice. *Cognit. Sci.* 42, 1777–1832. doi: 10.1111/cogs.12626

Codreanu, E., Sommerhoff, E., Huber, S., Ufer, S., and Seidel, T. (2020). Between authenticity and cognitive demand: finding a balance in designing a video-based simulation in the context of mathematics teacher education. *Teach. Teacher Educat.* 95:103146. doi: 10.1016/j.tate.2020.103146

Cooksey, R. W., Freebody, P., and Wyatt-Smith, C. (2007). Assessment as judgment-in-context: analysing how teachers evaluate students' writing. *Educat. Res. Evaluat.* 13, 401–434. doi: 10.1080/13803610701728311

Cooksey, R. W., Freeboy, P., and Davidson, G. R. (1986). Teachers' predictions of children's early reading achievement: an application of social judgment theory. *Am. Educat. Res. J.* 23, 41–64. doi: 10.3102/00028312023001041

Corno, L. (2008). On teaching adaptively. *Educat. Psychol.* 43, 161–173. doi: 10.1080/00461520802178466

Cortina, K. S., Miller, K. F., McKenzie, R., and Epstein, A. (2015). Where low and high inference data converge: validation of CLASS assessment of mathematics instruction using mobile eye tracking with expert and novice teachers. *Int. J. Sci. Math Educat.* 13, 389–403. doi: 10.1007/s10763-014-9610-5

Csanadi, A., Eagan, B., Kollar, I., Shaffer, D. W., and Fischer, F. (2018). When coding-and-counting is not enough: using epistemic network analysis (ENA) to analyze verbal data in CSCL research. *Int. J. Comput. Supported Collaborative Learn.* 13, 419–438. doi: 10.1007/s11412-018-9292-z

DeAngelus, M., and Pelz, J. B. (2009). Top-down control of eye movements: yarbus revisited. *Visual Cognition* 17, 790–811. doi: 10.1080/13506280902793843

Deary, I. J., Strand, S., Smith, P., and Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence* 35, 13–21. doi: 10.1016/j.intell.2006.02.001

Förster, N., and Böhmer, I. (2017). "Das Linsenmodell - Grundlagen und exemplarische Anwendungen in der pädagogisch-psychologischen Diagnostik [The lens-model—basics and exemplary applications in educational-psychological diagnostics]," in *Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen*, eds A. Südkamp and A.-K. Praetorius (Münster, New York: Waxmann), 46–50.

Fredricks, J. A., Blumenfeld, P. C., and Paris, A. H. (2004). School engagement: potential of the concept, state of the evidence. *Rev. Educat. Res.* 74, 59–109. doi: 10.3102/00346543074001059

Fredricks, J. A., Filsecker, M., and Lawson, M. A. (2016a). Student engagement, context, and adjustment: addressing definitional, measurement, and methodological issues. *Learn. Instruct.* 43, 1–4. doi: 10.1016/j.learninstruc.2016.02.002

Fredricks, J. A., Wang, M.-T., Schall Linn, J., Hofkens, T. L., Sung, H., Parr, A., et al. (2016b). Using qualitative methods to develop a survey measure of math and science engagement. *Learn. Instruct.* 43, 5–15. doi: 10.1016/j.learninstruc.2016.01.009

Funder, D. C. (1995). On the accuracy of personality judgment: a realistic approach. *Psychol. Rev.* 102, 652–670. doi: 10.1037/0033-295X.102.4.652

Funder, D. C. (2012). Accurate personality judgment. *Curricul. Direct. Psychol. Sci.* 21, 177–182. doi: 10.1177/0963721412445309

Garcia, E. B., Sulik, M. J., and Obradović, J. (2019). Teachers' perceptions of students' executive functions: disparities by gender, ethnicity, and ELL status. *J. Educat. Psychol.* 111, 918–931. doi: 10.1037/edu0000308

Gegenfurtner, A. (2020). *Professional Vision and Visual Expertise*. Regensburg: University of Regensburg.

Gegenfurtner, A., Lehtinen, E., and Säljö, R. (2011). Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. *Educat. Psychol. Rev.* 23, 523–552. doi: 10.1007/s10648-011-9174-7

Glock, S., Krolak-Schwerdt, S., Klapproth, F., and Böhmer, M. (2012). Improving teachers' judgments: accountability affects teachers' tracking decision. *Int. J. Tech. Inclusive Educat.* 1, 89–98. doi: 10.20533/ijtie.2047.0533.2012.0012

Glock, S., Krolak-Schwerdt, S., Klapproth, F., and Böhmer, M. (2013). Beyond judgment bias: how students' ethnicity and academic profile consistency influence teachers' tracking judgments. *Soc. Psychol. Educat.* 16, 555–573. doi: 10.1007/s11218-013-9227-5

Goodwin, C. (1994). Professional vision. *Am. Anthropolog.* 96, 606–633. doi: 10.1525/aa.1994.96.3.02a00100

Grossman, P., Compton, C., Igra, D., Romfeldt, M., Shahan, E., and Willimanson, P. W. (2009). Teaching practice: a cross-professional perspective. *Teachers College Record* 111, 2055–2100.

Haataja, E., Garcia Moreno-Esteva, E., Salonen, V., Laine, A., Toivanen, M., and Hannula, M. S. (2019). Teacher's visual attention when scaffolding collaborative mathematical problem solving. *Teach. Teach. Educat.* 86:102877. doi: 10.1016/j.tate.2019.102877

Haataja, E., Salonen, V., Laine, A., Toivanen, M., and Hannula, M. S. (2020). The relation between teacher-student eye contact and teachers' interpersonal behavior during group work: a multiple-person gaze-tracking case study in secondary mathematics education. *Educat. Psychol. Rev.* 8:229. doi: 10.1007/s10648-020-09538-w

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Ufer, S., et al. (2019). Facilitating diagnostic competence in simulations: a conceptual framework and a research agenda for medical and teacher education. *Frontline Learn. Res.* 7, 1–24. doi: 10.14786/flr.v7i4.384

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., et al. (2018). Teachers' assessment competence: integrating knowledge-, process, and product-oriented approaches into a competence-oriented conceptual model. *Teach. Teacher Educat.* 76, 181–193. doi: 10.1016/j.tate.2017.12.001

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., and Van de Weijer, J. (2011). *Eye Tracking: A Comprehensive Guide to Methods and Measures.* First edition. Oxford, New York, NY; Auckland: Oxford University Press.

Hörstermann, T., Pit-ten Cate, I. M., Krolak-Schwerdt, S., and Glock, S. (2017). "Primacy effects in attention, recall and judgment patterns of simultaneously presented student information: Evidence from an eye-tracking study," in *Student Achievement: Perspectives, Assessment and Improvement Strategies*, ed G. Hughes (Hauppauge: Nova Science Publishers Incorporated), 1–28.

Huang, C. (2011). Self-concept and academic achievement: a meta-analysis of longitudinal relations. *J. School Psychol.* 49, 505–528. doi: 10.1016/j.jsp.2011.07.001

Huber, S. A., Häusler, J., Jurik, V., and Seidel, T. (2015). Self-underestimating students in physics instruction: development over a school year and its connection to internal learning processes. *Learn. Individ. Diff.* 43, 83–91. doi: 10.1016/j.lindif.2015.08.021

Huber, S. A., and Seidel, T. (2018). Comparing teacher and student perspectives on the interplay of cognitive and motivational-affective student characteristics. *PLoS ONE* 13:e0200609. doi: 10.1371/journal.pone.0200609

Iacobucci, D., Posavac, S. S., Kardes, F. R., Schneider, M. J., and Popovich, D. L. (2015a). The median split: robust, refined, and revived. *J. Consumer Psychol.* 25, 690–704. doi: 10.1016/j.jcps.2015.06.014

Iacobucci, D., Posavac, S. S., Kardes, F. R., Schneider, M. J., and Popovich, D. L. (2015b). Toward a more nuanced understanding of the statistical properties of a median split. *J. Consumer Psychol.* 25, 652–665. doi: 10.1016/j.jcps.2014.12.002

Jacobs, V. R., Lamb, L. L., and Philipp, R. (2010). Professional noticing of children's mathematical thinking. *J. Res. Mathemat. Educat.* 41, 169–202. doi: 10.5951/jresematheduc.41.2.0169

Jansen, M., Lüdtke, O., and Schroeders, U. (2016). Evidence for a positive relation between interest and achievement: examining between-person and within-person variation in five domains. *Contemporary Educat. Psychol.* 46, 116–127. doi: 10.1016/j.cedpsych.2016.05.004

Jarodzka, H., Balsev, T., Homqvist, K., Nyström, M., Scheiter, K., Gerjets, P., et al. (2012). Conveying clinical reasoning based on visual observation via eye-movement modelling examples. *Instruct. Sci.* 5, 813–827. doi: 10.1007/s11251-012-9218-5

Jarodzka, H., Holmqvist, K., and Gruber, H. (2017). Eye tracking in educational science: theoretical frameworks and research agendas. *J. Eye Movement Res.* 10, 1–18. doi: 10.16910/jemr.10.1.3

Jarodzka, H., van Gog, T., Dorr, M., Scheiter, K., and Gerjets, P. (2013). Learning to see: guiding students' attention via a model's eye movements fosters learning. *Learn. Instruct.* 25, 62–70. doi: 10.1016/j.learninstruc.2012.11.004

Jurik, V., Gröschner, A., and Seidel, T. (2013). How student characteristics affect girls' and boys' verbal engagement in physics instruction. *Learn. Instruct.* 23, 33–42. doi: 10.1016/j.learninstruc.2012.09.002

Jurik, V., Gröschner, A., and Seidel, T. (2014). Predicting students' cognitive learning activity and intrinsic learning motivation: how powerful are teacher statements, student profiles, and gender? *Learn. Individ. Diff.* 32, 132–139. doi: 10.1016/j.lindif.2014.01.005

Kaiser, J., Retelsdorf, J., Südkamp, A., and Möller, J. (2013). Achievement and engagement: how student characteristics influence teacher judgments. *Learn. Instruct.* 28, 73–84. doi: 10.1016/j.learninstruc.2013.06.001

Karing, C. (2009). Diagnostische Kompetenz von Grundschul- und Gymnasiallehrkräften im Leistungsbereich und im Bereich Interessen [Diagnostic competence of primary and secondary school teachers regarding achievement and interest]. *Zeitschrift für Pädagogische Psychol.* 23, 197–209. doi: 10.1024/1010-0652.23.34.197

Karst, K., and Bonefeld, M. (2020). Judgment accuracy of preservice teachers regarding student performance: the influence of attention allocation. *Teach. Teacher Educ.* 94:103099. doi: 10.1016/j.tate.2020.103099

Kaufmann, E., Reips, U.-D., and Wittmann, W. W. (2013). A critical meta-analysis of lens model studies in human judgment and decision-making. *PLoS ONE* 8:e83528. doi: 10.1371/journal.pone.0083528

Keppens, K., Consuegra, E., Goossens, M., Maeyer, S., and de Vanderlinde, R. (2019). Measuring pre-service teachers' professional vision of inclusive classrooms: a video-based comparative judgement instrument. *Teach. Teach. Educat.* 78, 1–14. doi: 10.1016/j.tate.2018.10.007

Kersting, N. B., Sutton, T., Kalinec-Craig, C., Stoehr, K. J., Heshmati, S., Lozano, G., et al. (2016). Further exploration of the classroom video analysis (CVA) instrument as a measure of usable knowledge for teaching mathematics: taking a knowledge system perspective. *ZDM Mathematics Educat.* 48, 97–109. doi: 10.1007/s11858-015-0733-0

Kim, L. E., and Klassen, R. M. (2018). Teachers' cognitive processing of complex school-based scenarios: differences across experience levels. *Teach. Teach. Educat.* 73, 215–226. doi: 10.1016/j.tate.2018.04.006

Krauzlis, R. J., Goffart, L., and Hafed, Z. M. (2017). Neuronal control of fixation and fixational eye movements. *Philosophic. Transact. R. Soc. London* 372:20160205. doi: 10.1098/rstb.2016.0205

Krolak-Schwerdt, S., Böhmer, M., and Gräsel, C. (2013). The impact of accountability on teachers' assessments of student performance: a social cognitive analysis. *Soc. Psychol. Educat.* 16, 215–239. doi: 10.1007/s11218-013-9215-9

Kuncel, N. R., Klieger, D. M., Connelly, B. S., and Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: a meta-analysis. *J. Appl. Psychol.* 98, 1060–1072. doi: 10.1037/a0034156

Lachner, A., Jarodzka, H., and Nückles, M. (2016). What makes an expert teacher? Investigating teachers' professional vision and discourse abilities. *Instruct. Sci.* 44, 197–203. doi: 10.1007/s11251-016-9376-y

Lam, S.-,f., Jimerson, S., Shin, H., Cefai, C., Veiga, F. H., Hatzichristou, C., et al. (2016). Cultural universality and specificity of student engagement in school: the results of an international study from 12 countries. *Br. J. Educat. Psychol.* 86, 137–153. doi: 10.1111/bjep.12079

Lau, S., and Roeser, R. W. (2008). Cognitive abilities and motivational processes in science achievement and engagement: a person-centered analysis. *Learn. Individual Diff.* 18, 497–504. doi: 10.1016/j.lindif.2007.11.002

Linnenbrink-Garcia, L., Pugh, K. J., Koskey, K. L. K., and Stewart, V. C. (2012). Developing conceptual understanding of natural selection: the role of interest, efficacy, and basic prior knowledge. *J. Exp. Educat.* 80, 45–68. doi: 10.1080/00220973.2011.559491

Loibl, K., Leuders, T., and Dörfler, T. (2020). A framework for explaining teachers' diagnostic judgements by cognitive modeling (DiaCoM). *Teach. Teach. Educat.* 91:103059. doi: 10.1016/j.tate.2020.103059

Machts, N., Kaiser, J., Schmidt, F. T. C., and Möller, J. (2016). Accuracy of teachers' judgments of students' cognitive abilities: a meta-analysis. *Educat. Res. Rev.* 19, 85–103. doi: 10.1016/j.edurev.2016.06.003

Marksteiner, T., Reinhard, M.-A., Dickhäuser, O., and Sporer, S. L. (2012). How do teachers perceive cheating students? Beliefs about cues to deception and detection accuracy in the educational field. *Eur. J. Psychol. Educat.* 27, 329–350. doi: 10.1007/s10212-011-0074-5

Marquart, C. L., Hinojosa, C., Swiecki, Z., Eagan, B., and Shaffer, D. W. (2018). *Epistemic Network Analysis (Version 1.7.0) [Software]*. Available online at: http://app.epistemicnetwork.org

Marsh, H. W., and Martin, A. J. (2011). Academic self-concept and academic achievement: relations and causal ordering. *Br. J. Educat. Psychol.* 81, 59–77. doi: 10.1348/000709910X503501

McIntyre, N. A., and Foulsham, T. (2018). Scanpath analysis of expertise and culture in teacher gaze in real-world classrooms. *Instruct. Sci.* 46, 435–455. doi: 10.1007/s11251-017-9445-x

McIntyre, N. A., Jarodzka, H., and Klassen, R. M. (2019). Capturing teacher priorities: using real-world eye-tracking to investigate expert teacher priorities across two cultures. *Learn. Instruct.* 60, 215–224. doi: 10.1016/j.learninstruc.2017.12.003

McIntyre, N. A., Mainhard, M. T., and Klassen, R. M. (2017). Are you looking to teach? Cultural, temporal and dynamic insights into expert teacher gaze. *Learn. Instruct.* 49, 41–53. doi: 10.1016/j.learninstruc.2016.12.005

Meissel, K., Meyer, F., Yao, E. S., and Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: exploring student characteristics that influence teacher judgments of student ability. *Teach. Teach. Educat.* 65, 48–60. doi: 10.1016/j.tate.2017.02.021

Meschede, N., Fiebranz, A., Möller, K., and Steffensky, M. (2017). Teachers' professional vision, pedagogical content knowledge and beliefs: on its relation and differences between student and in-service teachers. *Teach. Teach. Educat.* 66, 158–170. doi: 10.1016/j.tate.2017.04.010

Nestler, S., and Back, M. D. (2013). Applications and extensions of the lens model to understand interpersonal judgments at zero acquaintance. *Curr. Direct. Psychol. Sci.* 22, 374–379. doi: 10.1177/0963721413486148

Nyström, M., Andersson, R., Holmqvist, K., and van de Weijer, J. (2013). The influence of calibration method and eye physiology on eyetracking data quality. *Behav. Res. Methods* 45, 272–288. doi: 10.3758/s13428-012-0247-4

Oudman, S., van de Pol, J., Bakker, A., Moerbeek, M., and van Gog, T. (2018). Effects of different cue types on the accuracy of primary school teachers' judgments of students' mathematical understanding. *Teach. Teach. Educat.* 76, 214–226. doi: 10.1016/j.tate.2018.02.007

Praetorius, A.-K., Berner, V.-D., Zeinz, H., Scheunpflug, A., and Dresel, M. (2013). Judgment confidence and judgment accuracy of teachers in judging self-concepts of students. *J. Educat. Res.* 106, 64–76. doi: 10.1080/00220671.2012.667010

Praetorius, A.-K., Drexler, K., Rösch, L., Christophel, E., Heyne, N., Scheunpflug, A., et al. (2015). Judging students' self-concepts within 30s? Investigating judgement accuracy in a zero-acquaintance situation. *Learn. Individ. Diff.* 37, 231–236. doi: 10.1016/j.lindif.2014.11.015

Praetorius, A.-K., Koch, T., Scheunpflug, A., Zeinz, H., and Dresel, M. (2017). Identifying determinants of teachers' judgment (in)accuracy regarding students' school-related motivations using a Bayesian cross-classified multi-level model. *Learn. Instruct.* 52, 148–160. doi: 10.1016/j.learninstruc.2017.06.003

Rimm-Kaufman, S. E., Baroody, A. E., Larsen, R. A. A., Curby, T. W., and Abry, T. (2015). To what extent do teacher–student interaction quality and student gender contribute to fifth graders' engagement in mathematics learning? *J. Educat. Psychol.* 107, 170–185. doi: 10.1037/a0037252

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., and Spinath, F. M. (2015). Intelligence and school grades: a meta-analysis. *Intelligence* 53, 118–137. doi: 10.1016/j.intell.2015.09.002

Sabers, D. S., Cushing, K. S., and Berliner, D. C. (1991). Differences among teachers in a task characterized by simultaneity, multidimensional, and immediacy. *Am. Educat. Res. J.* 28, 63–88. doi: 10.3102/00028312028001063

Santagata, R., and Yeh, C. (2016). The role of perception, interpretation, and decision making in the development of beginning teachers' competence. *ZDM Math. Educat.* 48, 153–165. doi: 10.1007/s11858-015-0737-9

Schiefele, U., Krapp, A., and Winteler, A. (1992). "Interest as a predictor of academic achievement: a meta-analysis research," in *The Role of Interest in Learning and Development*, eds K. A. Renninger, S. Hidi, and A. Krapp (Hillsdale: Erlbaum), 183–212.

Schnitzler, K., Holzberger, D., and Seidel, T. (2020). All better than being disengaged: student engagement patterns and their relations to academic self-concept and achievement. *Eur. J. Psychol. Educat.* doi: 10.1007/s10212-020-00500-6. [Epub ahead of print].

Schütz, A. C., Braun, D. I., and Gegenfurtner, K. R. (2011). Eye movements and perception: a selective review. *J. Vision* 11:9. doi: 10.1167/11.5.9

Seidel, T. (2006). The role of student characteristics in studying micro teaching–learning environments. *Learn. Environ. Res.* 9, 253–271. doi: 10.1007/s10984-006-9012-x

Seidel, T., Jurik, V., Häusler, J., and Stubben, S. (2016). "Mikro-Umwelten im Klassenverband: Wie sich kognitive und motivational-affektive Schülervoraussetzungen auf die Wahrnehmung und das Verhalten im Fachunterricht auswirken [Micro-environments in the classroom: How cognitive and motivational-affective student characteristics affect perception and behavior in class]," in *Bedingungen und Effekte guten Unterrichts*, eds N. McElvany, W. Bos, H. G. Holtappels, M. M. Gebauer, and F. Schwabe (Münster: Waxmann Verlag), 65–87.

Seidel, T., Schnitzler, K., Kosel, C., Stürmer, K., and Holzberger, D. (2020). Student characteristics in the eyes of teachers: differences between novice and expert teachers in judgment accuracy, observed behavioral cues, and gaze. *Educat. Psychol. Rev.* 15:98. doi: 10.1007/s10648-020-09532-2

Seidel, T., and Stürmer, K. (2014). Modeling and measuring the structure of professional vision in preservice teachers. *Am. Educat. Res. J.* 51, 739–771. doi: 10.3102/0002831214531321

Shaffer, D. W., Collier, W., and Ruis, A. R. (2016). A tutorial on epistemic network analysis: analyzing the structure of connections in cognitive, social, and interaction data. *J. Learn. Analyt.* 3, 9–45. doi: 10.18608/jla.2016.33.3

Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., et al. (2009). Epistemic network analysis: a prototype for 21st-century assessment of learning. *Int. J. Learn. Media* 1, 33–53. doi: 10.1162/ijlm.2009.0013

Shaffer, D. W., and Ruis, A. R. (2017). "Epistemic network analysis: A worked example of theory-based learning analytics," in *Handbook of Learning Analytics*, eds C. Lang, G. Siemens, A. Wise, and D. Gašević. First edition (Solar Society for Learning Analytics Research), 175–187. doi: 10.18608/hla17.015

Shavelson, R. J., Hubner, J. J., and Stanton, G. C. (1976). Self-concept: validation of construct interpretations. *Rev. Educat. Res.* 46, 407–441. doi: 10.3102/00346543046003407

Sherin, M. G., and van Es, E. A. (2009). Effects of video club participation on teachers' professional vision. *J. Teacher Educat.* 60, 20–37. doi: 10.1177/0022487108328155

Shulman, L. S. (1986). Those who understand: knowledge growth in teaching. *Educat. Res.* 15, 4–14. doi: 10.3102/0013189X015002004

Shulman, L. S. (1987). Knowledge and teaching: foundations of the new reform. *Harvard Educat. Rev.* 57, 1–21. doi: 10.17763/haer.57.1.j463w79r56455411

Sinatra, G. M., Heddy, B. C., and Lombardi, D. (2015). The challenges of defining and measuring student engagement in science. *Educat. Psychol.* 50, 1–13. doi: 10.1080/00461520.2014.1002924

SMI (2017a). *BeGaze: SMI SensoMotoric Instruments.*

SMI (2017b). *Experiment Center: SMI SensoMotoric Instruments.*

Snow, R. E. (1989). "Cognitive-conative aptitude interactions in learning," in *Abilities, Motivation, and Methodology: The Minnesota Symposium on Learning and Individual Differences*, eds R. Kanfer, P. L. Ackerman, and R. Cudeck (New York, NY: Lawrence Erlbaum Associates), 435–474. doi: 10.4324/9780203762905

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz [Accuracy of teachers' assessment of student characteristics and the construct of diagnostic competence]. *Zeitschrift für Pädagogische Psychologie* 19, 85–95. doi: 10.1024/1010-0652.19.12.85

Star, J. R., and Strickland, S. K. (2008). Learning to observe: using video to improve preservice mathematics teachers' ability to notice. *J. Math. Teach. Educat.* 11, 107–125. doi: 10.1007/s10857-007-9063-7

Steinmayr, R., and Spinath, B. (2009). The importance of motivation as a predictor of school achievement. *Learn. Individual Diff.* 19, 80–90. doi: 10.1016/j.lindif.2008.05.004

Stürmer, K., Könings, K. D., and Seidel, T. (2013). Declarative knowledge and professional vision in teacher education: effect of courses in teaching and learning. *Br. J. Educat. Psychol.* 83, 467–483. doi: 10.1111/j.2044-8279.2012.02075.x

Stürmer, K., Seidel, T., and Holzberger, D. (2016). Intra-individual differences in developing professional vision: preservice teachers' changes in the course of an innovative teacher education program. *Instruct. Sci.* 44, 293–309. doi: 10.1007/s11251-016-9373-1

Stürmer, K., Seidel, T., Müller, K., Häusler, J., and Cortina, K. (2017). What is in the eye of preservice teachers while instructing? An eye-tracking study about attention processes in different teaching situations. *Zeitschrift für Erziehungswissenschaft* 20, 75–92. doi: 10.1007/s11618-017-0731-9

Südkamp, A., Kaiser, J., and Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: a meta-analysis. *J. Educat. Psychol.* 104, 743–762. doi: 10.1037/a0027627

Südkamp, A., Praetorius, A.-K., and Spinath, B. (2018). Teachers' judgment accuracy concerning consistent and inconsistent student profiles. *Teach. Teach. Educat.* 76, 201–213. doi: 10.1016/j.tate.2017.09.016

Thiede, K. W., Brendefur, J. L., Osguthorpe, R. D., Carney, M. B., Bremner, A., Strother, S., et al. (2015). Can teachers accurately predict student performance? *Teach. Teach. Educat.* 49, 36–44. doi: 10.1016/j.tate.2015.01.012

Urhahne, D., and Zhu, M. (2015). Accuracy of teachers' judgments of students' subjective well-being. *Learn. Individ. Diff.* 43, 226–232. doi: 10.1016/j.lindif.2015.08.007

Valentine, J. C., DuBois, D. L., and Cooper, H. (2004). The relation between self-beliefs and academic achievement: a meta-analytic review. *Educat. Psychol.* 39, 111–133. doi: 10.1207/s15326985ep3902_3

van den Bogert, N., van Bruggen, J., Kostons, D., and Jochems, W. (2014). First steps into understanding teachers' visual perception of classroom events. *Teach. Teach. Educat.* 37, 208–216. doi: 10.1016/j.tate.2013.09.001

van Es, E. A., and Sherin, M. G. (2002). Learning to notice: scaffolding new teachers' interpretations of classroom interactions. *J. Tech. Teach. Educat.* 10, 571–596.

van Es, E. A., and Sherin, M. G. (2008). Mathematics teachers' "learning to notice" in the context of a video club. *Teach. Teach. Educat.* 24, 244–276. doi: 10.1016/j.tate.2006.11.005

Wolff, C. E., Jarodzka, H., van den Bogert, N., and Boshuizen, H. P. A. (2016). Teacher vision: expert and novice teachers' perception of problematic classroom management scenes. *Instruct. Sci.* 44, 243–265. doi: 10.1007/s11251-016-9367-z

Wooldridge, A. R., Carayon, P., Shaffer, D. W., and Eagan, B. (2018). Quantifying the qualitative with epistemic network analysis: a human factors case study of task-allocation communication in a primary care team. *IISE Transact. Healthcare Syst. Eng.* 8, 72–82. doi: 10.1080/24725579.2017.1418769

Wyss, C., Rosenberger, K., and Bührer, W. (2020). Student teachers' and teacher educators' professional vision: findings from an eye tracking study. *Educat. Psychol. Rev.* 43:175. doi: 10.1007/s10648-020-09535-z

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

frontiers in Education

BRIEF RESEARCH REPORT
published: 25 January 2021
doi: 10.3389/feduc.2020.584200

Formative Decision-Making in Response to Primary Science Classroom Assessment: What to do Next?

*Sarah Earle**

School of Education, Bath Spa University, Bath, United Kingdom

Classroom assessment is purposeful when the information is utilised by teachers to support learning. Such formative assessment practices can be difficult to enact in a primary science classroom, with the whole class often involved in practical activities and with limited lesson time. This preliminary study seeks to explore formative decision-making and the subsequent actions taken by teachers in the classroom. Primary teachers who used a Teacher Assessment in Primary Science (TAPS) Focused Assessment activity were asked to describe what action they took as a result of the classroom interactions stimulated by the activity. 142 teachers in 9 regions of England completed a paper questionnaire at a training day. The qualitative data pertinent to the study was extracted and thematic content analysis carried out to determine the kinds of actions and changes to practice that were described. It was found that the "next step" described by teachers varied in timing; some made changes within the lesson, others provided follow up activities or made longer-term adaptation to teaching practices. Being responsive to the assessment information provided by the children took many forms, for example, supporting pupils to reflect on investigations during the lesson, discussing vocabulary or concepts, providing time for further exploration, or explicit modeling of science skills. Formative decisions were taken at a whole class level, rather than making individual adaptations. It is argued that enabling teachers to be more explicit about their tacit decision-making could support them to make more formative use of assessment information to support pupil learning.

Keywords: TAPS, design-based research, primary science, formative assessment, teacher assessment literacy

OPEN ACCESS

Edited by:
Gavin T. L. Brown,
The University of Auckland,
New Zealand

Reviewed by:
Wei Shin Leong,
Ministry of Education, Singapore
Hui Yong Tay,
Nanyang Technological University,
Singapore

***Correspondence:**
Sarah Earle
s.earle@bathspa.ac.uk

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

Received: *16 July 2020*
Accepted: *21 December 2020*
Published: *25 January 2021*

Citation:
Earle S (2021) Formative Decision-Making in Response to Primary Science Classroom Assessment: What to do Next?. Front. Educ. 5:584200. doi: 10.3389/feduc.2020.584200

INTRODUCTION

Formative assessment has widely been hailed as key to supporting children's learning (e.g. Black and William, 1998; Harlen, 2013; Wiliam, 2018). Gardner et al. (2010) assert that assessment should focus on improving learning, explaining why formative assessment became commonly known as "Assessment for Learning" (Assessment Reform Group, 1999). However, in practice, formative use of assessment information has been difficult to implement, with changes to teacher practice taking time and often skewed by current policy such as an increased focus across the world on using assessment for accountability (DeLuca et al., 2019). The difficulties encountered with implementation indicate that there is still a need for further research in this area, with the aim of finding manageable ways for teachers to make use of formative assessment in the classroom. Low levels of teacher assessment literacy or capability across the profession (Gardner, 2007) also point to a

need for further support for teachers. This article aims to explore the teacher decision-making involved in utilising formative assessment information, in order make such processes more explicit and enhance understanding in the field.

Assessment education is in a constant state of flux (DeLuca et al., 2019) as it responds to changes in assessment policy and emphasis. Significant changes have taken place in the last decade in England, with primary schools (for children aged 4–11) following a new statutory National Curriculum with assessment indicators presented as age-related expectations (Department for Education, 2013). Science is included as a core subject, but with English and mathematics featuring on school league tables, primary science is often perceived to be of lower status, meaning less time for both teaching and teacher professional learning (CFE Research, 2017). Such time pressures mean that primary science lessons would typically cover a different topic each lesson, with it being normal to "move on" once the lesson was taught. This means that there would be little time for follow up or extension discussions, little time for acting upon formative assessment information.

Formative assessment, with its focus on supporting learning, could be a useful tool for schools dealing with the impact of the Covid-19 global pandemic. Modeling from seasonal learning research suggests that attainment may slow or decline during long periods of school closures (Kuheld and Tarasawa, 2020). Others have suggested that disadvantaged children are more likely to experience a such a "learning loss," further widening the gap between children from lower socio-economic backgrounds and their more affluent peers (Education Endowment Foundation, 2020; Müller and Goldenberg, 2020). However, an over-emphasis on "identifying gaps" on a return to school may miss the point of formative assessment. Focusing on "lost learning" via frequent testing has long been identified as a misinterpretation of formative assessment (Klenowski, 2009); identifying the "gap" is only a precursor to formative action.

The purpose of formative assessment is to inform decisions about future learning experiences in the classroom (Harlen, 2007). Strategies associated with formative assessment include: identifying and making explicit success criteria; elicitation of children's existing ideas; feedback; self-assessment and peer assessment (Wiliam, 2018). However, these strategies are not separate to classroom teaching, formative assessment is embedded, it is part of the teaching process. For researchers, this makes it difficult to monitor, but for teachers, this means it should not add to their workload. By following such an approach, any interaction with pupils can provide useful assessment information. Such "assessment interactions" point to the need for planning and teaching to be responsive rather than wholly decided in advance: the interaction becomes formative when it provokes a response, when "action" is taken. Black and Wiliam suggest that: *"assessment provides information to be used as feedback . . . Such assessment becomes "formative assessment" when the evidence is actually used to adapt the teaching work to meet the needs"* (1998: 2), thus it is the use of assessment information to support the learning process which distinguishes formative and summative assessment, rather than the assessment task itself. Use of such assessment information could include: judgment according to criteria or comparison with previous performance in similar events to identify ongoing areas of concern, consideration of next steps, decision making and then formative action. Such formative assessment interactions and actions can be at the class, group or individual level.

Webb and Jones (2009) note that change in teacher practice is difficult and takes time, with practice needing time to be trialled, integrated and embedded. Teacher assessment literacy is a developmental process that requires teacher's reflection and critical evaluation of their diverse use assessment (Deluca et al., 2016). Assessment literacy or capability also requires an understanding of the subject being taught: content and pedagogical content knowledge (PCK, Shulman, 1986), since the teacher needs an understanding of the subject matter to be able to make judgements regarding pupil understanding, as well as pedagogical understanding of the most appropriate ways to teach and assess the content. Assessment capability cannot be separated from the subject context (Edwards, 2013). The teacher needs knowledge of the key concepts to identify what to assess and knowledge of assessment processes to identify how to assess and what to do with the information gained. This means that professional learning around assessment needs subject-specific elements for it to be usable in the classroom.

The Primary Science Teaching Trust funds the Teacher Assessment in Primary Science (TAPS) project (2013+) to develop support for teachers, which includes a range of examples and activity plans on the TAPS website linked to each of the curricular in the four countries of the United Kingdom (TAPS Website, 2020). TAPS uses a Design-Based Research methodology, which promotes collaboration between teachers and researchers, involving iterative cycles to trial and refine both resources and theoretical principles to impact educational practice (Design-Based Research Collective, 2003; Anderson and Shattuck, 2012; Easterday et al., 2018). The principles of formative assessment provide the theoretical basis for guidance to support teacher decision-making (Davies et al., 2017). When applied to the primary science classroom, these principles emphasise the elicitation of pupil understanding, the responsiveness of teachers to adapt their lessons in response to this information from the pupils and the active role of pupils in self and peer assessment (Wiliam, 2018).

One strand of TAPS, which is still evolving in the iterative cycles, is the use of a Focused Assessment approach for teaching and assessing scientific inquiry (Davies and McMahon, 2003). This approach proposes that one element of inquiry becomes the focus for teacher attention and any pupil drawing or writing, within the context of a whole inquiry. For example, in an investigation dropping different sized paper "spinners" (or helicopters) the teacher selects one part which will be given more teaching time. For example, a focus on recording results could include time on drawing tables or graphs; a focus on controlling variables could include more time planning and setting up the investigation; whilst a focus on drawing conclusions could involve individual writing to draw conclusions from the results. Selection of a focus in this way is designed to make teaching and assessment more manageable in a practical lesson. The TAPS Focused Assessment activity plans are

being trialled in all four countries of the United Kingdom, and the approach has become the subject of a large randomised control trial across England, which is being funded by the Education Endowment Foundation and the Wellcome Trust.

The TAPS Focused Assessment approach provides practical guidance for suggested activities, but carrying out the activities is not the same as implementing formative assessment. Formative assessment requires action, something needs to be done with the information gained from interactions with pupils. This study sought to find out what teachers who have carried out a TAPS task do next, whether they use the assessment information to tweak their teaching, what kind of action is taken and when this takes place. This study analyses initial findings to answer the following research question:

RQ. How do primary teachers act on information arising from classroom interactions stimulated by the TAPS Focused Assessment activities?

METHODS

This preliminary study of teacher-decision making is a small part of the larger TAPS project, which utilises a Design-Based Research methodology of iterative and collaborative research and development (Anderson and Shattuck, 2012). In order to answer the RQ, teachers undertaking TAPS training were directly asked to describe their practice; a purposive sample (Teddlie and Yu, 2007) who would be able to comment on their classroom interactions in response to TAPS Focused Assessment activities.

During the 2019–20 academic year, 142 teachers in 9 regions across England took part in TAPS Focused Assessment professional development (first day in October 2019, second day in January/February 2020, third day canceled due to Covid-19). The training included explanation of the science inquiry process (since many primary school teachers are not science specialists), together with consideration of assessment strategies. In between training days, the teachers were asked to carry out some TAPS Focused Assessment activities with their class and then feedback about their use at the next training day. On Day 2, teachers discussed their experiences, shared pupil work and completed a paper questionnaire (sharing of further cycles of formative assessment did not take place due to the cancellation of Day 3). As part of the questionnaire, all teachers were asked to describe the activities carried out with their class and what they did as a result of such classroom interactions. Responses to the question "What did you do next?" form the basis for this study. The teachers were explicitly asked to provide details of the *changes* to their practice. Such changes indicate formative use of information: teachers changing their practice in response to information gained from interactions with the children.

An open-ended question was selected so that teachers described their practice rather than assigned it to a pre-determined category (Oppenheim, 1992), particularly important for such a preliminary study to find out how teachers acted on the classroom information. It should be noted that the lead trainer was also the lead researcher, which may have influenced the teacher responses, however, it also enabled a fuller understanding of the teacher responses for this preliminary study. The study is qualitative, exploring the participant experience, but the size of the sample does enable numerical summaries for discussion of prevalence. The sample consisted of half science subject leaders (teaching any year group) and half Year five teachers (pupils aged 9–10). For full teacher details, see **Table 1**.

TABLE 1 | Age groups taught by teachers in the sample

Year group	Pupil age in years	No. of teachers in sample
R	4–5	3
Y1	5–6	5
Y2	6–7	6
Mixed key stage 1	4–7	5
Y3	7–8	9
Y4	8–9	12
Mixed lower key stage 2	7–9	8
Y5	9–10	73
Y6	10–11	9
Mixed upper key stage 2	9–11	12
	Total	**142**

In line with ethical procedures, all teachers were fully informed regarding the collection, use and storage of their questionnaire answers. They were also given the opportunity to withdraw their data (BERA, 2018). The paper questionnaires were anonymised at the point of typing up and then stored securely.

The data for the "What did you do next?" question was extracted into a spreadsheet for this study. Thematic content analysis was carried out on the 142 descriptions. They were sorted thematically into emergent groups and this was revisited multiple times to ensure that the final themes represented the dataset. Initially the data was sorted twice: in terms of timing of the described "next step" or action (during the lesson, extensions to the lesson, future teaching) and separately into the type of action described (changes to the teaching, the next tasks given to the children, the children's groups etc). Types of teacher action mapped onto the timings for when this took place, with different kinds of action happening at different time, for example, children's groups could be changed in the following lesson, but this did not happen during the same lesson. Thus the final themes presented below consist of types of action, placed into time order.

RESULTS

The majority of teachers described an action, something that they did next in response to information gained from interactions with pupils during the TAPS lesson. With time pressured primary science, the "normal" next step would be to move on to the next topic as per the pre-written school planning, so taking an action which extended or adapted the lesson for example, would indicate that the teachers were making a formative decision.

Thematic groups emerged in terms of both the kind of action taken and whether the action took place immediately: as part of the same lesson; soon after in a follow up lesson; or the adaptation

TABLE 2 | Frequency of formative actions described by teachers.

Timing of described formative action	Type of formative action described	Frequency in sample
1. As part of the same lesson or activity	a. More time on pupil recording of investigation e.g. draw diagram, take/annotate photos, complete table/graph, write prediction/conclusion etc	12
	b. More time for discussion with pupils of results, conclusions, evaluations or reflections	19
2. Activity follow up (extended or next lesson on the same topic)	a. Repeat or extend same investigation or lesson	10
	b. New focus on vocabulary	15
	c. New focus on concepts	17
	d. New investigation	12
3. Adaptation of future teaching (later that term)	a. Change to grouping of children in class	5
	b. Explicit teaching or practice of science skill	18
	c. Modeling or use of examples	8
	d. New focus on mathematics skill	4
	e. For teacher's own knowledge or future use	5
4. To support other teachers	(formative use in subject leader role rather than class teacher role)	5
5. No description of action	a. Answer did not include action	4
	b. No response to question	8
	Total	142

of future teaching. Examples of the thematic groups are listed below, following a frequency table to show prevalence of the teacher actions in **Table 2**.

Theme 1a. As part of same lesson or activity, the teacher's next step focused on a change to pupil recording of the results, such as drawing a table or graph (N = 12). For example, using a "planning board" from the TAPS resources to support children to construct a graph, or the addition of discussion time for children to discuss how they had recorded their results:

"Children worked as a group to put their data onto a bar graph - we used the planning boards to help to know where to put the correct data" (Teacher 13).

"After spinners - discussion explaining how to record our results" (Teacher 87).

Theme 1b. As part of same lesson or activity, the teacher noted the discussion of conclusions, perhaps supporting the pupils to evaluate or reflect on their investigation (N = 19). In a normal primary science lesson, full discussion of conclusions is often difficult to include because it needs to take place at the end of the lesson, when "tidying up" time may seemingly take priority, so making time for this "review" stage of the investigation indicates a change from normal practice. For example:

"Discussed the results with the class and got those who understood to share their findings with others" (Teacher 18).

"Reflection on their own planning of an experiment using their recordings/findings, how would they replan/do differently" (Teacher 5).

"Review - what did we find out? How could it change?" (Teacher 97).

Theme 2a. In a follow up to the activity, which could take place in an immediate lesson extension (continuing the same lesson) or continue into the next lesson (on the same topic), the teacher may support the pupils to repeat or extend their investigation (N = 10). Finding time for this (and the other actions below) indicates a change to the normal practice of moving on to the next topic. For example:

"*Allowed them/us time to carry out improvements. Gave them time to record" (Teacher 17).*

"Let the pupils choose other materials to test" (Teacher 15).

Theme 2b. Other actions following the activity focused on the pupils' use of vocabulary (N = 15). For example:

"*Identify areas of weakness to build on e.g. including scientific vocab in conclusions/explanations" (Teacher 26).*

"Ensure vocabulary displayed in classroom/table mats. Quick quiz at the start of a lesson to recap vocabulary and ensure retention" (Teacher 33).

"Give children opportunities to explore and discuss scientific vocabulary more in depth before investigation and sharing their interpretations" (Teacher 45).

Theme 2c. For some teachers, the next step involved further consideration of conceptual understanding (N = 17). For example:

"Verbal recap of different forces. Explaining what each force does" (Teacher 25).

"I showed them videos of a harp - real life example of pitch with different lengths of string" (Teacher 32).

"Returned to asking questions about air resistance - concept cartoons and post-it notes to elicit" (Teacher 42).

"Discussed misconceptions as a class. Discussed particles and why types of sugar dissolve certain ways" (Teacher 110).

Theme 2d. Other teachers chose to continue the investigating the same topic or set up new inquiries on the same topic (N = 12), rather than moving on to the next topic, which would have been normal practice. For example:

"Set up further experiments based on reversing dissolving" (Teacher 80).

"Next we investigated shadows so we used the planning booklets and post-its more confidently but only recorded the table and results in books." They found it hard to look at patterns in data so did some discreet work on this from "Handling Science Data Y5." From this we then did Biscuit Dunk to compare our Y4 and 6 reflections to look at progression of skills (Teacher 131).

"Applied results of investigation to real-life scenarios - new footpath at school" (Teacher 133).

Theme 3a. For some teachers, the formative assessment information was not used immediately, it was used to adapt future lessons, like in changing the way pupils were grouped (N = 5). For example:

"Grouped children differently for follow up lesson for insulation lesson" (Teacher 1).

"In the groups, assigned a role to each child e.g. to time, to measure, to record, etc. Discussed conclusions, improvements, etc as a class" (Teacher 75).

Theme 3b. In response to formative assessment information, some teachers decided to adapt their future lessons by being more explicit in their teaching of science skills (N = 18). For example:

"I taught how to identify variables and make choices" (Teacher 9).

"Lots of work around fair testing - use of planning grids with post-its as a whole class (only changing one variable). Use of planning grids in small groups - scaffolded at each stage" (Teacher 85).

"Changed to have a measure focus and taught each child to read from the scale" (Teacher 134).

Theme 3c. Some teachers planned to make more use of modeling and examples in their next lessons (N = 8). For example:

"Give examples of ways we could measure and discuss which was most appropriate" (Teacher 21).

"I modeled to children how to write a conclusion. Next experiment - children wrote their own conclusion" (Teacher 86).

Theme 3d. For other teachers, they decided mathematics skills needed a focus in future lessons (N = 4). It is not clear whether these were changes to planned maths lessons, but this indicates a recognition from the teacher of the interplay between the subjects. For example:

"Maths - thermometer lesson. 1 key recorder (whiteboard)" (Teacher 30).

"In maths - looked at graphs - in particular line graphs (scales)" (Teacher 35).

Theme 3e. A small number of teachers described how they would use the experience to feed into later teaching, but without a specific next step (N = 5). Such lack of specific action could indicate a lack of clarity or use of the formative assessment information. For example:

"Used lesson to plan for the next input" (Teacher 74).

"Follow the plan more and focus on individual children to ascertain learning" (Teacher 99).

Theme 4. For those teachers with subject leadership roles, the next step was more about supporting other teachers, rather than specific next steps for pupils (N = 5). For example:

"Meet with the staff to moderate progress across the school. Talk about and note down next steps" (Teacher 108).

"Reflected with staff" (Teacher 136).

Theme 5. A small number of teachers did not answer the question (N = 8) or described the interaction with pupils, without explaining how the information would be utilised (N = 4). This included activities carried out at the end of term, or teachers who were trialling the activities with pupils who they did not normally teach. For example:

"This was the final lesson at the end of term" (Teacher 59).

"The graphs were marked but (they were not my class) there were no formative comments.

The students did not follow up this activity with either conclusions or evaluations" (Teacher 139).

It is important to note that a "what next" question requires an end point to an activity, which may not take into account the ongoing responsive teaching taking place. It should also be noted that the action, or planned action, described by the teachers is specific to the activity. It would be expected that the same teacher may take a different action in a different situation, since responsive teaching is context-specific. Tracking the effectiveness of such feedforward actions was not possible in this study, both due to Covid-19 school closures and the difficulties of following the effect of individual formative actions without access to the classroom setting. The impact of teachers' formative actions on children's learning, in the short, medium and long term, merits further research.

DISCUSSION

The focus for this article was to explore teacher formative decision-making, to consider the kind of actions teachers took as a result of assessment interactions. The majority of teachers in this sample described an action, a change to practice, using the information gained from an interaction with pupils to make a decision about what to do next with the class. Making such changes, to for example adapt the lesson end or subsequent lesson, suggests that teachers were using the assessment information formatively (Black and Wiliam, 1998).

Findings from this preliminary study suggest that teachers may adapt their practice: within the lesson (Theme 1), when following up the lesson (Theme 2) or when planning future teaching (Theme 3); each of which will be considered in turn.

The Theme 1 actions "within the lesson" included the addition of discussions with children regarding pupil recording of results (12 teachers) or drawing conclusions (19 teachers). It could be questioned whether the teachers interpreted the question "what happened next" to mean "what happened after the practical activity?" and so just described the end of the lesson. However, during the second training day both of the Theme 1 actions were raised by teachers, for example, discussing how they had made more time for reflecting on investigation findings, so their responses could indicate that these were key areas that had changed in their practice. Pupil recording (drawing, writing etc) was discussed a number of times at the training day, both considering how to record the results of a science investigation in terms of the layout of a table or graph to help with interpretation of results, and the bigger question of how much of an investigation the pupils should be writing down. Traditional experiment write ups (Method, Results, Conclusion) can take a large amount of time for younger children, meaning that the lessons can become focused on the mechanics of writing rather than science content. The essence of the TAPS Focused

Assessment approach is that an inquiry focus should be selected for the lesson, for example, if conclusions is the focus then more lesson time is devoted to this and it is this part of the inquiry which the pupils will record. By spending more time on areas which may normally be neglected (e.g. with lesson time often running out before discussing conclusions), the teachers could be applying their training to both act on formative assessment interaction and maintain a science focus for the lesson.

Many teachers identified elements that merited further class time on the same topic, for example, extending the investigation, and addressing gaps in vocabulary or conceptual understanding (Theme 2). Such decisions require flexibility in planning and teaching time, together with appropriate content knowledge regarding the choice of which are the key concepts to follow up (Shulman, 1986; Edwards, 2013). It is important to note that these were class-wide interventions. The teacher had identified something that a number of children in the class struggled with, so this could be considered formative assessment at a class level rather than an individual level. This is perhaps an indication that use of formative assessment information needs to be manageable, especially in a subject that is only taught once a week. There cannot be an expectation of individual interventions for each of the pupils in a class of 30, but if many of them struggle with one element, a "tipping point" is reached and it is a reasonable adjustment to change or adapt subsequent sessions to address this. Looking at formative assessment at a class level could lead to individual needs going unaddressed, but teaching is a balance of what is ideal and what is possible. This suggests that in practical primary science lessons, it may often only be manageable to gauge a general level of understanding or performance, and individual assessments may need a different elicitation strategy.

For those teachers who described adapting future teaching, there was consideration of explicit teaching of skills (science or mathematics) with modeling of further examples to support pupils' learning (Theme 3). Again this occurred largely at the class level, with future planning adapted to include more opportunities to teach and practice. Identification of next steps for learning is a key feature of formative assessment. It appears that the majority of teachers in this sample were able to use the information gained from the assessment interactions to decide on next steps for learning. Whether such knowledge is acted upon immediately within the lesson, or in subsequent lessons, depends on a variety of factors about the lesson content and the class. It is not necessarily preferable to act immediately, since it may be that content needs revisiting in a different way or with a different example. It also may not be possible to act immediately, if areas for development are not identified until the end or after the lesson. Nevertheless, if teachers only "log" areas for development, but never get the chance to return to them, then formative assessment does not fulfill its purpose.

This study has found that teacher decision-making in response to formative assessment interactions can result in changes in the same or future lessons. Teaching adaptations include making space in the lesson schedule for further discussion and reflection, or explicit teaching and modeling of particular skills or concepts. It was found that in the time pressured primary science context, balancing the need to support individual learning with whole class manageability may lead to formative decision-making which is a "best fit" approach for the class. Decision-making within lesson time is difficult, especially when the teacher is busy managing the classroom activities as well as collecting assessment information. Such decision-making can be supported by subject-specific assessment training, since teachers need both assessment literacy and subject knowledge to be able to consider and decide on appropriate next steps. Developing an appropriate classroom assessment language to articulate and share evidence, decisions and the effectiveness of future actions with colleagues could lead to more assessment capable teachers (DeLuca et al., 2019). Enabling teachers to be more explicit about their tacit decision-making, could support them to make more formative use of assessment information to support pupil learning. This study was only able to consider one cycle of teacher action and reflection due to Covid-19 school closures, so further research is needed to explore changes to teacher practice over time and whether such formative decision-making processes become embedded in teacher practice or are just a one off "project effect". Research into such professional learning is ongoing, with the TAPS project working with teachers across the United Kingdom to collaboratively design principled support for the use of formative assessment in primary science.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article can be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Bath Spa University. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

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

FUNDING

The Teacher Assessment in Primary Science (TAPS) project benefits from funding from the Primary Science Teaching Trust, the Education Endowment Foundation and the Wellcome Trust.

ACKNOWLEDGMENTS

Thanks goes to the teachers and members of the TAPS team who share their thoughts about assessment so readily.

REFERENCES

Anderson, T., and Shattuck, J. (2012). Design-based research: a decade of progress in education research?. *Educ. Res.* 41 (1), 16–25. doi:10.3102/0013189x11428813

Assessment Reform Group (1999). *Assessment for learning: beyond the black box*. Cambridge: University of Cambridge Faculty of Education.

BERA (2018). *Ethical Guidelines for Educational Research*. 4th Ed. London: BERA.

Black, P., and Wiliam, D. (1998). *Inside the Black Box*. London: GL Assessment.

CFE Research (2017). *State of the nation' report of UK primary science education: baseline research for the Wellcome Trust Primary Science Campaign*. Leicester: CFE Research.

Davies, D., Earle, S., McMahon, K., Howe, A., and Collier, C. (2017). Development and exemplification of a model for teacher assessment in primary science. *Int. J. Sci. Educ.* 39 (14), 1869–1890. doi:10.1080/09500693.2017.1356942

Davies, D., and McMahon, K. (2003). Assessment for enquiry: supporting teaching and learning in primary science. *Sci. Educ. Int.* 14 (4), 29–39. doi:10.1039/D0RP00283F

DeLuca, C., LaPointe-McEwan, D., and Luhanga, U. (2016). Approaches to Classroom Assessment Inventory: a new instrument to support teacher assessment literacy. *Educ. Assess.* 21 (4), 248–266. doi:10.1080/10627197.2016.1236677

DeLuca, C., Willis, J., Cowie, B., Harrison, C., Coombs, A., Gibson, A., , et al.Trask, S. (2019). Policies, programs and practices: exploring the complex dynamics of assessment education in teacher education across four countries. *Front. Educ.* 4, 132. doi:10.3389/feduc.2019.00132

Department for Education (DfE) (2013). *National Curriculum in England: science programmes of study*. London: DfE.

Design-Based Research Collective (2003). Design-based research: an emerging paradigm for educational inquiry. *Educ. Res.* 32 (No. 1), p5–8. doi:10.3102/0013189X032001005

Easterday, M., Rees Lewis, D., and Gerber, E. (2018). The logic of design research. *Learn. Res. Pract.* 4 (2), 131–160. doi:10.1080/23735082.2017.1286367

Education Endowment Foundation (2020). Impact of school closures on the attainment gap: rapid evidence assessment. https://educationendowmentfoundation.org.uk/covid-19-resources/best-evidence-on-impact-of-school-closures-on-the-attainment-gap Accessed 3 June 2020.

Edwards, F. (2013). Quality assessment by science teachers: five focus areas. *Sci. Educ. Int.* 24 (2), 212–226. doi:10.1039/D0RP00121J

Gardner, J., Harlen, W., Hayward, L., and Stobartwith Montgomery M, G. (2010). *Developing teacher assessment*. Maidenhead: Oxford University Press.

Gardner, J. (2007). Is teaching a 'partial' profession? *Making the Grade*, *Summer*, 28, 18–21.

Harlen, W. (2013). *Assessment and inquiry-based science education: issues in policy and practice*. Trieste: Global Network of Science Academies.

Harlen, W. (2007). *Assessment of learning*. London: Sage.

Klenowski, V. (2009). Assessment for learning revisited: an asia-pacific perspective. *Assess Educ. Princ. Pol. Pract.* 16 (3), 263–268. doi:10.1080/09695940903319646

Kuhfeld, M., and Tarasawa, B. (2020). The COVID-19 slide: What summer learning loss can tell us about the potential impact of school closures on student academic achievement. *Collaborative for Student Growth: Brief* https://www.nwea.org/blog/2018/summer-learning-loss-what-we-know-what-were-learning Accessed June 2, 2020.

Müller, L. M., and Goldenberg, G. (2020). Education in times of crisis: the potential implications of school closures for teachers and students https://chartered.college/2020/05/07/chartered-college-publishes-report-into-potential-implications-of-school-closures-and-global-approaches-to-education Accessed June 2, 2020.

Oppenheim, A. (1992). *Questionnaire design, interviewing and attitude measurement*. 2nd edition. London: Pinter Publishers.

Shulman, L. (1986). Those who understand: knowledge growth in teaching. *Educ. Res.* 15 (2), 4–14. doi:10.3102/0013189x015002004

Teacher Assessment in Primary Science (TAPS) Website (2020). https://pstt.org.uk/resources/curriculum-materials/assessment Accessed July 15, 2020.

Teddlie, C., and Yu, F. (2007). Mixed methods sampling: a typology with examples. *J. Mix. Methods Res.* 1, 77–100. doi:10.1177/1558689806292430

Webb, M., and Jones, J. (2009). Exploring tensions in developing assessment for learning, *Assessment in Education: Principles. Policy Pract.* 16 (2), 165–184. doi:10.1080/09695940903075925

Wiliam, D. (2018). *Embedded formative assessment*. 2nd ed. Bloomington: Solution Tree Press.

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.

frontiers in Education

ORIGINAL RESEARCH
published: 26 February 2021
doi: 10.3389/feduc.2021.604568

Selecting Mathematical Tasks for Assessing Student's Understanding: Pre-Service Teachers' Sensitivity to and Adaptive Use of Diagnostic Task Potential in Simulated Diagnostic One-To-One Interviews

Stephanie Kron, Daniel Sommerhoff, Maike Achtner and Stefan Ufer*

Chair of Mathematics Education, LMU Munich, Germany

OPEN ACCESS

Edited by:
Dennis Alonzo, University of New South Wales, Australia

Reviewed by:
Eric C K Cheng, The Education University of Hong Kong, Hong Kong
Wei Shin, Leong, Ministry of Education, Singapor

****Correspondence:***
Stephanie Kron
kron@math.lmu.de

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

Received: *09 September 2020*
Accepted: *06 January 2021*
Published: *26 February 2021*

Citation:
Kron S, Sommerhoff D, Achtner M and Ufer S (2021) Selecting Mathematical Tasks for Assessing Student's Understanding: Pre-Service Teachers' Sensitivity to and Adaptive Use of Diagnostic Task Potential in Simulated Diagnostic One-To-One Interviews. Front. Educ. 6:604568. doi: 10.3389/feduc.2021.604568

Teachers' diagnostic competences are regarded as highly important for classroom assessment and teacher decision making. Prior conceptualizations of diagnostic competences as judgement accuracy have been extended to include a wider understanding of what constitutes a diagnosis; novel models of teachers' diagnostic competences explicitly include the diagnostic process as the core of diagnosing. In this context, domain-general and mathematics-specific research emphasizes the importance of tasks used to elicit student cognition. However, the role of (mathematical) tasks in diagnostic processes has not yet attracted much systematic empirical research interest. In particular, it is currently unclear whether teachers consider diagnostic task potential when selecting tasks for diagnostic interviews and how this relationship is shaped by their professional knowledge. This study focuses on pre-service mathematics teachers' selection of tasks during one-to-one diagnostic interviews in live simulations. Each participant worked on two 30 mins interviews in the role of a teacher, diagnosing a student's mathematical understanding of decimal fractions. The participants' professional knowledge was measured afterward. Trained assistants played simulated students, who portrayed one of four student case profiles, each having different mathematical (mis-) conceptions of decimal fractions. For the interview, participants could select tasks from a set of 45 tasks with different diagnostic task potentials. Two aspects of task selection during the diagnostic processes were analyzed: participants' sensitivity to the diagnostic potential, which was reflected in higher odds for selecting tasks with high potential than tasks with low potential, and the adaptive use of diagnostic task potential, which was reflected in task selection influenced by a task's diagnostic potential in combination with previously collected information about the student's understanding. The results show that participants vary in their sensitivity to diagnostic task potential, but not in their adaptive use. Moreover, participants' content knowledge had a significant effect on their sensitivity. However, the effects of pedagogical content and pedagogical knowledge did not reach significance. The results highlight that pre-service teachers require further support to effectively attend to diagnostic task potential. Simulations were used for assessment

purposes in this study, and they appear promising for this purpose because they allow for the creation of authentic yet controlled situations.

Keywords: diagnostic competences, student assessment, diagnostic task potential, teacher education, professional knowledge, teacher decision making

INTRODUCTION

Teachers' pedagogical decisions are contingent on reliable assessments of students' understanding (Van de Pol et al., 2010). Recently, teachers' competences with regard to diagnosing student understanding have attracted increasing research focus, (e.g. Herppich et al., 2018; Leuders et al., 2018). Previous research has established a relationship between teachers' judgment accuracy and student learning (Behrmann and Souvignier, 2013). The paradigm of judgment accuracy conceptualizes diagnostic competences as the match between teachers' expectations of individual students' test performance and the students' actual test performance (Südkamp et al., 2012). The judgment accuracy paradigm has often been criticized for only considering diagnosis in the form of an estimated test score and for not investigating the diagnostic process itself (Südkamp & Praetorius, 2017; Herppich et al., 2018). Extending the concept of judgment accuracy has triggered more comprehensive approaches toward diagnostic competences (Praetorius et al., 2012; Aufschnaiter et al., 2015), which include a wider understanding of what constitutes a diagnosis, as well as the diagnostic process itself. While the first extension led to the inclusion of students' (mis-)conceptions, understanding, and strategies for diagnosing (Herppich et al., 2018), the second extension targets how teachers actually collect information to form their diagnostic judgement. The current study focuses on the second extension, specifically the diagnostic process.

Heitzmann et al. (2019) define diagnosing as the goal-directed accumulation and integration of information to reduce uncertainty when making educational decisions. Generating diagnostic information about students' understanding requires some form of teacher-student interaction (Klug et al., 2013), which may take place in very different situations (Karst et al., 2017). Diagnostic interviews with individual students about specific mathematical concepts have been highlighted as a prototypical example of such situations, (e.g. Wollring, 2004). While teachers may also diagnose "on the fly" during whole-group classroom discussions, or while supporting individual or small-group work, these situations allow for the detailed study of diagnostic processes and their disentanglement from subsequent pedagogical decisions (Kaiser et al., 2017).

Diagnosing requires that teachers have diagnostic competences, which are conceptualized as individual, cognitive, and context-sensitive dispositions (Koeppen et al., 2008; Ufer & Neumann, 2018) that become observable through the accuracy of the diagnostic processes and the subsequent diagnoses. Diagnostic competences, in this sense, enable "[. . .] people to apply their knowledge in diagnostic activities according to professional standards to collect and interpret data in order to make high-quality decisions" (Heitzmann et al., 2019). Tröbst et al. (2018) emphasize the importance of teachers' professional knowledge to diagnostic competence. The lack of tools to investigate, measure, and foster (pre-service) teachers' diagnostic competences (Südkamp et al., 2008; Praetorius et al., 2012) highlights the importance of simulations, which provide an authentic environment to investigate diagnostic competences under controlled conditions, as well as the possibility of improving participants' skills based on the generated findings. Klug et al. (2013) point out that teachers mainly assess their students during face-to-face interaction in the classroom. Reconstructing such situations in simulations is discussed as a promising approach to apply newly learned knowledge in authentic situations, especially in the context of pre-service teacher education (Grossman et al., 2009).

As mentioned, prior works emphasize the need to understand the diagnostic process as the link between individual dispositions, such as professional knowledge, and the quality of diagnostic judgements and subsequent decisions (Heitzmann et al., 2019). This process includes diverse activities, such as the elicitation of diagnostic information from students, the observation and interpretation of the resulting student answers, and the integration of these interpretations into a diagnostic judgement that facilitates valid pedagogical decision making (Herppich et al., 2018; Heitzmann et al., 2019; Loibl et al., 2020). In this contribution, we focus on how teachers use tasks to elicit diagnostic information. We propose two constructs characterizing promising task selection during the diagnostic process: sensitivity *to* the diagnostic potential of tasks and adaptive use of diagnostic task potential. We introduce operationalizations of these constructs in an authentic diagnostic simulation and analyze the kinds of professional knowledge that underlie these aspects of task selection.

Process Models of Teachers' Diagnoses

Existing models of the diagnostic process usually try to cover a wide range of diagnostic situations, for example, ranging from formal to informal assessment or from formative to summative assessment, or assessment based on verbal interaction *vs*. written documents (Philipp, 2018). The field has moved from generic models, for example, Klug, et al (2013) model, which closely resembles general self-regulation models, to more specific models that describe how diagnostic information is gathered and processed. For example, the NeDiKo model (Herppich et al., 2018) describes diagnostic processes in teachers' professional practice as a sequence of prototypical decisions and subsequent diagnostic actions. It particularly focuses on accumulating diagnostic information by generating hypotheses and testing them with data collected from students. Similarly, the COSIMA model (Heitzmann et al., 2019) describes this process as

an orchestration of eight diagnostic activities, including generating hypotheses, generating diagnostic evidence, evaluating diagnostic evidence, and drawing conclusions from this evidence. In contrast, the DiaKom model (Loibl et al., 2020) explicitly distinguishes observable situation characteristics from latent person characteristics and cognitive processes, and conceptualizes cognitive processes along the PID model (Blömeke et al., 2015) as perceiving features from a situation, interpreting them against professional knowledge, and making pedagogical decisions.

Except for the DiaKom model (which does not explicitly include observable actions), each model describes teacher actions that are intended to elicit diagnostically relevant information from students (evidence generation in the COSIMA model). The models are open to a wide range of possible evidence generation methods, such as administering a standardized test, making informal observations during class, and analyzing students' work on exams or homework. In most of these cases, eliciting diagnostically relevant information involves assigning a student some kind of task (orally or in written form) and interpreting the student's answers or responses. Studies based on the DiaKom model, (e.g. Ostermann et al., 2018) usually do not include such explicit evidence generation actions by teachers; rather, they provide participants with prepared diagnostic information. However, diagnosis in professional practice will, to a large extent, be initiated and coordinated by the teacher. Thus, it is a vital question how and based on what knowledge teachers select tasks to elicit and accumulate diagnostically relevant information about students' understanding of particular concepts.

The Role of Professional Knowledge in Diagnosing

As described, professional knowledge is assumed to be a central individual resource underlying diagnostic competences, and is indeed part of most models of diagnostic competences. In the COSIMA model, it is listed as one of the central resources (Heitzmann et al., 2019); in the NeDiKo model (Herppich et al., 2018), it is subsumed under "dispositions"; and in the DiaKom model (Loibl et al., 2020), it is part of teachers' person characteristics. In all of these models, the influence of teachers' professional knowledge on their diagnoses, judgments, and decisions is mediated by the characteristics of the diagnostic process. Accordingly, how teachers apply their professional knowledge during a diagnostic situation is assumed to be the main link between their individual knowledge and the final diagnosis (cf. Blömeke et al., 2015).

While different conceptualizations of professional knowledge by knowledge type or content can be adopted (Förtsch et al., 2018), the categorization of Schulman (1987) is central in the context of teacher education. He proposes structuring teachers' professional knowledge (among others) into content knowledge (CK), pedagogical content knowledge (PCK), and pedagogical knowledge (PK). CK describes knowledge about the subject matter, which, in our case, is mathematics. In a manner similar to the mathematical knowledge for teaching (MKT) measures (Ball et al., 2008) many studies have conceptualized CK as sound knowledge of school mathematics and its background (COACTIV: Baumert & Kunter 2013; TEDS-M: Blömeke et al., 2011; KiL: Kleickmann et al., 2014). PCK is necessary specifically for teaching a specific subject and usually consists of knowledge about student cognition, knowledge about instructional and diagnostic tasks, and knowledge about instructional approaches and strategies (Baumert and Kunter, 2013). PK refers to knowledge about teaching and learning in general, with no connection to a specific subject (Kleickmann et al., 2014).

Südkamp et al. (2012) indicates that CK and PCK do have an influence on teachers' judgment accuracy. Beyond this, several studies have shown that components of teachers' professional knowledge predict their diagnostic competence. For example, Van den Kieboom et al. (2013) investigated pre-service teachers' questioning during one-to-one diagnostic interviews. They found that participants with low CK only affirmed students' responses, without asking any probing questions. Participants with higher CK asked probing questions to investigate students' understanding and try to help improve it.

Ostermann et al. (2018) investigated the influence of PCK on pre-service teachers' estimates of task difficulty in an intervention study. Participants who received PCK instruction produced more accurate estimates of task difficulty than participants from a control group. However, Herppich et al. (2018) point out that previous research findings do not allow for final conclusions to be drawn about the relationship between a person's professional knowledge and the given diagnosis.

Prior research is less clear about the role and interplay of the three components of professional knowledge and how these knowledge components are used in the diagnostic process. Herppich et al. (2018), for example, subsume all three components into teachers' assessment competences but do not describe their specific roles. Tröbst et al. (2018) point out that it is not yet known how the different components of professional knowledge interact in the diagnostic process. Against this background, empirical evidence about the relationship between the components of teachers' professional knowledge and the well-described characteristics of the diagnostic process is needed. To achieve this, we propose to study task selection as an important part of the diagnostic process, and analyze its connection to teachers' professional knowledge.

Task Selection in the Diagnostic Process

Neubrand et al. (2011) highlight the important role of tasks in mathematics instruction (Bromme, 1981) and further work has connected the quality of tasks (Baumert et al., 2010) and their implementation (Stein et al., 2008) with student learning. Black and Wiliam (2009) stress the use of learning tasks to elicit evidence of student understanding as a key strategy of formative assessment. Indeed, mathematical tasks not only play a role as learning opportunities for students, teachers also draw evidence about students' mathematical understanding from their responses to tasks (Schack et al., 2013). How teachers select and use tasks for diagnostic purposes, however, has attracted little attention in past research. To frame our approach to filling this

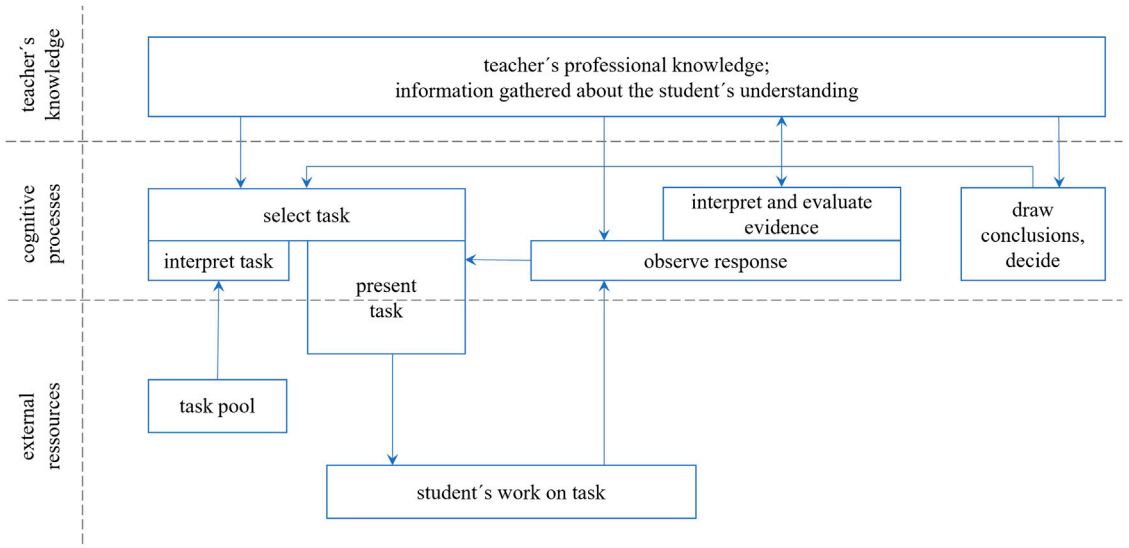


FIGURE 1 | Model of task-based diagnostic processes.

gap, we propose a process model of task-based diagnostic processes (**Figure 1**).

The core of the model consists of a set of latent cognitive processes (the middle portion), that draw on teachers' knowledge (the upper portion), as well as on external information and resources and visible actions (the lower portion). The arrows in the model indicate the flow of information between the (partially parallel) sub-processes and the other parts of the model. Regarding teachers' knowledge, we differentiate professional knowledge, including CK, PCK, and PK as well as temporary information about a specific student's mathematical understanding. This differentiation reflects claims from the literature that teachers' professional knowledge is, at least to some degree, situation-specific, (e.g. Lin and Rowland, 2016). Diagnosing draws on professional knowledge and aims to accumulate information about students' understanding. We assume an orchestration of three central cognitive processes to achieve this:

1) *Drawing conclusions and deciding*: Based on the available knowledge and accumulated information about the student's understanding, this process continuously monitors whether sufficient information is available to draw certain conclusions and make reliable decisions. It has a regulatory function, as missing information (to make a decision or judgment) influences the second process, task selection.
2) *Task selection* relies on the teacher's interpretation of the available tasks (as tools for evidence generation). Based on the accumulated information and evaluation in the previous process, it is vital to select tasks that will most likely add new evidence to the accumulated information. This process is strongly connected to the actual (observable) presentation of the task to the student, which usually triggers the student to show some kind of observable work on the task. Task presentation also includes asking follow-up questions based on student responses.
3) The third process is initially responsible for *observing and attending* to the student's work. Since perception is a knowledge-based and knowledge-driven process, it draws on professional knowledge and accumulated information (Sherin and Star, 2011). Based on this observation, another sub-process is responsible for *interpreting and evaluating* the evidence, for example, weighing more or less reliable parts of the observed evidence or integrating them into the accumulated information about the student's understanding.

Based on this perspective of task-based diagnosis, the selection of tasks becomes a crucial element of the diagnostic process. It has long been discussed that mathematical tasks can vary substantially regarding how much diagnostic information they can potentially unveil regarding, for example, a specific concept (Maier et al., 2010). A task that can be solved correctly with superficial strategies or even without understanding the concept has low potential to generate evidence about knowledge of this specific concept. For an example, the task of comparing the decimal fractions 0.417 and 0.3 has a low diagnostic potential regarding knowledge about decimal fractions, because even a student using superficial methods, (e.g. identifying 0.417 as the larger decimal fraction because 417 is larger than 3) could solve that task correctly. Without asking the student for further explanations, a teacher would not be able to generate reliable evidence as to whether the student can compare decimal fractions based on a proper understanding of place value principles. In the literature, the term "diagnostic potential of tasks" has been used, without an exact definition, to describe this dimension, and knowledge about the diagnostic potential of tasks has been repeatedly mentioned as part of mathematics teachers' professional knowledge, (e.g. Moyer-Packenham & Milewicz,

2002; Maier et al., 2010; Baumert & Kunter, 2013). We define the diagnostic potential of a task as its ability to stimulate student responses, allowing for the generation of reliable evidence about students' mathematical understanding. Asking a student to compare 0.354 to 0.55, for example, would have higher diagnostic potential than comparing 0.417 and 0.3 because the former task can only can be solved correctly when the student is capable of applying the underlying concept by comparing the values in each section. All known (systematic) superficial strategies of decimal comparison lead to incorrect decisions in this case.

The diagnostic potential of a mathematical task is primarily determined by the more or less elaborate mathematical strategies that can be used to solve the task. If only those strategies observed in typical student responses, which reflect at least some understanding of the underlying concepts, lead to a correct answer, this indicates high diagnostic potential. If the task can be solved correctly using one of the typically observed superficial strategies, which are not based on a reliable understanding of the underlying concepts, (e.g. treating the integer and the decimal part of the decimal number separately as if they were whole numbers), this implies low task potential.

In summary, we assume that the selection of tasks is an important part of diagnostic processes. It is reasonable to assume that task selection is not only influenced by teachers' knowledge and cognition, as well as the characteristics of the available tasks, but also by the characteristics of the student. In view of this, we propose to distinguish between two facets of teachers' selection of tasks during diagnosis: *sensitivity to* and *adaptive use of* diagnostic task potential.

Assuming that task selection is driven by information about the task and information about the student, we propose these two different facets of task selection. *Sensitivity to the diagnostic task potential* corresponds to the sub-processes "interpret task" and "select task". It becomes visible in the diagnostic potential of the tasks presented during the interview and relies on the characteristics of the task. *Adaptive use of the diagnostic task potential* corresponds to the arrow from "draw conclusions, decide" to "select task". It becomes visible in the extent to which a selected task can contribute new information about the student's understanding, beyond what has been collected before/could have been collected based on prior observations. This facet thus relies on task and student characteristics.

Sensitivity to Diagnostic Task Potential

We conceptualize sensitivity to the diagnostic potential of tasks as a facet of diagnostic competence that is observable in the diagnostic processes. Being sensitive to the diagnostic potential of tasks means considering a task's diagnostic potential to be an important factor during task selection. Participants' sensitivity to the diagnostic potential of tasks would be reflected in a higher probability of selecting tasks with high potential in comparison to tasks with low potential.

In line with existing models and first evidence on the role of professional knowledge in diagnosis, we assume that sensitivity to diagnostic task potential is related to teachers' professional knowledge (Baumert & Kunter, 2013). However, it is an open question as to which component of professional knowledge specifically underlies sensitivity to the diagnostic potential of tasks. Being able to select tasks with high potential requires that teachers be aware of and attend to relevant task characteristics in order to rate their diagnostic potential (e.g., Loibl et al., 2020). Baumert and Kunter (2013) conceptualize knowledge about the diagnostic potential of tasks as a part of PCK. However, other components may also play a role. For example, CK could be needed to identify the range of mathematical strategies that can be used to solve a task. PK might contribute to regulating the diagnostic process and, more specifically, to coordinating task selection in a superordinate manner.

Adaptive Use of Diagnostic Task Potential

A high level of importance may be attributed to selecting tasks with high diagnostic potential at the start of a diagnostic process when little or unreliable information about a student is available. However, once initial information has been gathered, this accumulated knowledge about learning characteristics should guide efficient diagnostic processes. Choosing an optimal task from a set of alternatives for a specific situation has been described under the term "adaptivity" in the past (Heinze & Verschaffel, 2009). We consider task selection adaptive (regarding the use of diagnostic potential) if the selected task can contribute evidence about facets of students' understanding beyond what was inferred from prior observations. In this sense, if the selection of a specific task is adaptive, it does not only depend on the characteristics of the task itself, but also on existing information about the specific student. Selecting tasks adaptively requires teachers to take accumulated information about student's mathematical understanding into account.

Adaptivity to the use of diagnostic task potential may lead to a different task selection than sensitivity to diagnostic potential alone. A task with low general diagnostic potential might offer additional information, as it may help to exclude a specific misconception that has not yet been considered. In contrast, a task with high general diagnostic potential might not offer additional information if it is redundant to what was visible in the tasks before. However, selecting tasks with high diagnostic potential (sensitivity) at the beginning of a diagnostic process might support the adaptive selection of diagnostic tasks later in the process, since more reliable information about students' understanding is available.

Deciding, whether the selection of a specific task is adaptive in a specific diagnostic situation might be almost impossible for an external observer, since neither the "real" understanding of a real student nor the information accumulated by the teacher on this student is observable. To allow for an empirical investigation, we thus propose an approximation to adaptivity. Task selection is adaptive, if a selected task can, in principle, deliver additional information about the student that goes beyond what could possibly have been observed in preceding tasks.

Regarding the role of components of teachers' professional knowledge for adaptivity regarding the use of diagnostic task potential, similar arguments can be made for sensitivity to diagnostic task potential. Adaptivity, however, puts a bigger demand on teachers' representation of the current information

about students' understanding than just being sensitive to the diagnostic task potential. We can assume that stronger PCK, for example, about student cognition, may support teachers in organizing, retaining, and utilizing this information. Even though this argument may specifically explain the connection between PCK and adaptivity, relationships with CK and PK may be expected based on our theoretical conceptualization.

THE CURRENT STUDY

Although the concept is frequently mentioned in the literature on teachers' professional knowledge and competences (Baumert & Kunter, 2013), it is not yet known how beginning and experienced teachers deal with the diagnostic potential of tasks when diagnosing students' understanding. In particular, no reliable evidence pertaining to teachers' actions in realistic situations is available. In the context of teacher education, it is crucial to understand how teachers can apply the knowledge, they acquired in university courses, to real-life situations. In this study, we investigate pre-service teachers' task-selection in authentic role-play simulations of diagnostic one-to-one interviews about decimal fractions. Each participant took part in two simulation sessions. In each session, a trained actor played one out of four pre-defined student case profiles. Each simulation consisted of two phases: an initial phase in which teachers could select from a restricted set of screening tasks and a second phase in which a larger set of diagnostic tasks was additionally available. Overall, the simulation-based approach allows for the control of factors related to the student whose understanding is being diagnosed, and thus provides a reliable measurement.

The main goal of this study is to introduce the constructs of sensitivity to the diagnostic potential of tasks and the adaptive use of this potential, and provide initial results pertaining to these characteristics of the diagnostic process for pre-service teachers. Moreover, we investigate whether there are systematic inter-individual differences in these process characteristics and how they relate to components of pre-service teachers' professional knowledge. To this end, the study focuses on the questions delineated in Sections Sensitive use of diagnostic task potential and Adaptive use of diagnostic task potential.

Sensitive Use of Diagnostic Task Potential

Our first goal was to obtain insights into whether pre-service teachers' task selection is sensitive to the varying diagnostic potential of tasks.

> *RQ1.1 To what extent are pre-service teachers sensitive to the diagnostic potential of tasks? Is there systematic variation in pre-service teachers' sensitivity to diagnostic potential?*

Based on the fact that participants in our study had already participated in a lecture and tutorials on mathematics education in the area of numbers and operations (including decimal fractions), we expected that they would show some sensitivity to diagnostic potential, that is participants choose tasks with high diagnostic potential with a higher probability than tasks with low diagnostic potential. We controlled for the interview position (first *vs.* second simulation), but expected small differences, at most, between the two interviews. Moreover, we predicted no significant differences in pre-service teachers' sensitivity over the four different student case profiles, but we expected systematic inter-individual variation between the participants in their tendency to prefer high-potential over low-potential tasks.

> *RQ1.2 To which extent is pre-service teachers' sensitivity to the diagnostic potential of tasks related to different components of their professional knowledge (CK, PCK, PK)?*

Based on the discussion emerging from prior research, we assumed that sensitivity to diagnostic potential would primarily be linked to the participants' PCK. Thus, we expected that higher PCK would go along with higher odds of choosing high potential tasks (over low-potential tasks).

Adaptive Use of Diagnostic Task Potential

Second, adaptive use of diagnostic task potential was investigated. To study adaptive use, only the second phase of each interview was considered. Based on the prior definition, task selection was considered adaptive if the selected task could provide additional evidence about a student's understanding beyond what could be observed in the initial (screening) phase of the interview; that is, the task has the diagnostic potential to yield information beyond what had already been gathered.

> *RQ2.1 To which extent is pre-service teachers' task selection adaptive to evidence generated from prior tasks? Is there systematic variation in pre-service teachers' adaptive use of diagnostic task potential?*

Even though adaptive use of the diagnostic potential of tasks can be considered a more complex demand than sensitivity to this potential, we expected pre-service teachers to show a higher probability of making task selections coded as adaptive beforehand compared with those coded as non-adaptive. Again, we expected only small differences across the two interview positions (first *vs.* second interview) and the four student case profiles. Regarding systematic variation in pre-service teachers' adaptive use of diagnostic task potential, we had no initial hypothesis, as research results on teachers' adaptivity are scarce, and we were not able to find relevant results in prior empirical mathematics education research.

> *RQ2.2 To which extent is pre-service teachers' adaptivity in the use of diagnostic task potential related to different components of their professional knowledge (CK, PCK, PK)?*

Following the assumption of sensitivity to diagnostic task potential, we assumed that the adaptive use of diagnostic task potential would also be related to the participants' PCK. Thus, we expected that higher PCK would go along with higher odds of making adaptive (*vs.* non-adaptive) task selections.

METHODS

To investigate these questions, we used data from simulated diagnostic one-to-one interviews about decimal fractions. In these role-play simulations, pre-service teachers engaged in diagnostic interviews with one of the four types of simulated students. The simulated students were played by teaching assistants who had been trained to enact the four different student case profiles. Each participant worked on the simulation twice, each time with a different student case profile.

Participants

The simulation was embedded in regular courses in pre-service teacher education at a large university in Germany. Every course participant was asked to perform the simulation. Participation in the study, including the provision of data for analysis, was voluntary and based on explicit consent. Performance during the simulation had no influence on course grades. Participation in the study was remunerated. The ethics committee and the data protection officer approved the study in advance.

The sample consisted of 65 pre-service high school teachers (38 f, 26 m, 1 day; M_{age} = 23.9, SD = 5.7). Most participants were in their fifth or lower semester. Almost all participants had finished at least four university courses focusing on mathematics education in previous semesters. Based on these degree programs' curricula, almost every participant should have completed one course focusing on the topic of decimal fractions before participating in the study. Two thirds of the participants reported, that they had taken at least one course specifically covering PCK on decimal fractions. Most participants had at least some experience teaching students from their practical studies; on average, they had held 9.15 lessons on their own (SD = 6.41). They had 2.0 years of experience in private tutoring, on average (SD = 1.57).

Procedure

Each participant took part in one half-day session. The simulation was held in a face-to-face setting, supported by a web-based interview system that guided the participants through the simulation.

After a short introduction explaining the goals and procedures of the sessions, the participants had 15 min to acquaint themselves with the interview system and the diagnostic tasks embedded in the system. They then met a trained teaching assistant who played the role of the simulated student. The participants had up to 30 min to select diagnostic tasks and pose them to the simulated student. The simulated student answered according to the applicable student case profile; responses were provided verbally or in writing. After the participants completed the interview, they had 15 min to compose a report containing their diagnosis of the student's understanding. Since the main focus of this contribution is the diagnostic process, the report phase will not be considered further. After the report on the first interview had been completed, the second interview started, following the same procedure as the first one (**Figure 2**).

After the participants had completed the two simulations, they were given a paper-and-pencil test (duration: 60 min) to assess their professional knowledge.

Design of the Role-Play-Based Live Simulation

The role-play-based live simulations were developed to investigate pre-service teachers' diagnostic competences (Marczynski et al., in press). These role-plays simulate a diagnostic interview between a mathematics teacher and a sixth-grade student. All the participants had an opportunity to play the teacher's role in the simulations. The implementation and design of the simulation received positive ratings from experts in a validation study (Stürmer et al., in press).

Based on prior research in mathematics education, the topic of decimal fractions was selected as the interview content because there is a substantial amount of research on students' understanding and misconceptions in this field (Steinle, 2004; Heckmann, 2006; Padberg & Wartha, 2017). To structure the interview, we distinguished the three central fields of knowledge about decimal fractions:

(1) Number representation in the decimal place value system, including comparison of decimals
(2) Basic arithmetic operations of addition, subtraction, multiplication, and division, including flexible and adaptive use of calculation strategies for the four basic arithmetic operations
(3) Connections between arithmetic operations and their meaning in realistic situations and word problems

Based on this framework, four student case profiles were constructed. These represented different profiles of sound understanding and misconceptions over these three fields of knowledge. Each student case profile has strong misconceptions in one of the three fields, partial misconceptions in a second field, and quite sound understanding in the remaining field.

To pre-structure the diagnostic interview, participants were given a set of 16 clusters of diagnostic tasks. The first three task clusters, with ten tasks in total, were designed as initial tasks, focusing on different aspects of knowledge about decimal fractions. The 13 subsequent task clusters, with 35 tasks in total, were designed to provide additional information based on what could have been observed in the initial tasks. The interview itself was separated into two phases. In the first phase of the interview, only the initial tasks were available to the participants. The subsequent tasks were unlocked after at least one task from each of the three initial task clusters was selected. There were no limitations on the number of selected tasks. Additionally, participants were allowed to create their own tasks using blank task templates, but this opportunity was used only rarely (in 11 out of 130 simulations).

Familiarization Phase

Before the one-to-one interviews started, the participants were introduced to their assignment during the simulation. They were asked to imagine being in the position of a high school teacher who offers consultation meetings for students struggling with mathematics learning. In these meetings, they should try to get an impression of the student's competences and misconceptions, in this case with regard

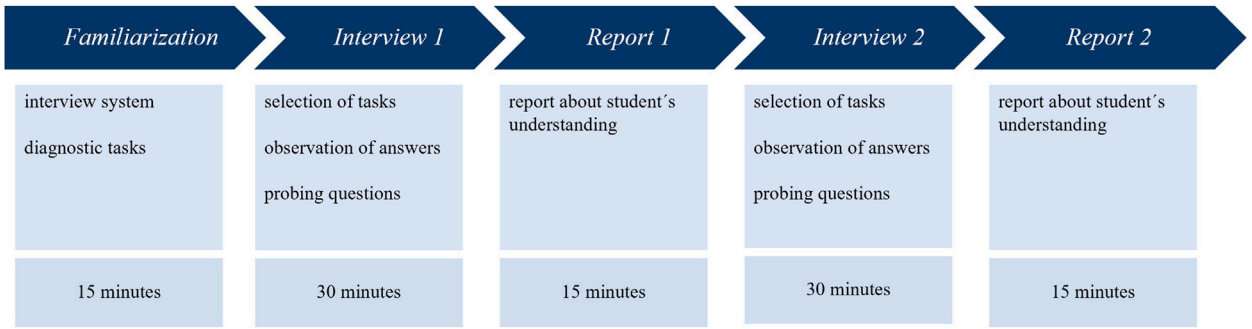


FIGURE 2 | Procedure of the simulation sessions.

to decimal fractions, by using a set of diagnostic tasks. They were informed that after the interview, they would be asked to write a report to the simulated student's teacher containing a post-interview diagnosis. Before the first simulated student joined the setting, participants had 15 min to acquaint themselves with the interview system and the diagnostic tasks that they could use to conduct the diagnostic interview. They were instructed to analyze the tasks with respect to their usefulness for generating evidence about students' mathematical understanding. The interview system offered a list of all available task clusters. Clicking on a cluster displayed a description and a list of the respective tasks. The participants were asked to make notes in the interview system during familiarization, and these notes were displayed during the interviews whenever the corresponding task was selected.

Interview Phase

After the familiarization phase, the simulated student joined the setting, and the first interview started. To diagnose the student's mathematical understanding, participants could select from the provided tasks. Clicking on a task cluster displayed a list similar to the one that appeared on participants' screen during familiarization together with the notes from the familiarization. The task itself was displayed on another screen, and the simulated student started to solve the task and write down the solution. The participants observed the student's response and asked for further explanations if needed. The participants could also make notes during the interview. As described above, only the first three clusters with the initial tasks were displayed at the beginning of the interview. Participants were instructed that they should choose at least one task from each of the three clusters and to use up to half of the interview for this first phase. After they had gathered enough information about the student's understanding or when the time limit was exceeded, they were moved on to the report phase.

Actor Training

The teaching assistants playing the students received standardized training spanning three half-day meetings. In the first meeting, they received theoretical background information about the content and aim of the study as well as their assignments during the simulations. For each of the four student case profiles, they received a detailed handout with background information about the relevant student's mathematical understanding and a handout with the student's handwritten solutions and verbal explanations for each task. The assistants were asked to familiarize themselves with the different student case profiles before the second meeting. In the second meeting, questions regarding the student case profiles were discussed, followed by a practical phase in which they were acquainted with the interview system. Subsequently, they worked on two simulations, once playing the student's role and once playing the teacher's role. Questions were discussed at the end of the meeting. In the final session, the assistants' command of each student case profile was tested in single standardized interviews.

Structure of the Dataset and Analytic Approach

The data in this study have a three-level structure. On the person level, the dataset contains participants' professional knowledge scores (person-parameters). Since each person worked on two simulations, data on the simulation level are nested within persons. At this level, the dataset contains the position of the simulation (first *vs.* second simulation) and which of the four student case profiles was used in the simulation. Finally, during each simulation, participants could decide whether to select each of the tasks from the diagnostic interview. These selections are nested in the simulations. The log files we extracted from the web-based interview system indicate whether a specific task was selected (or not) in a particular simulated interview.

Furthermore, we included two more variables, *diagnostic potential* and *adaptivity*, on the selection level. These were based on a priori coding of the tasks. Adaptivity was coded for each combination of task and student case profile. *Diagnostic potential* describes whether a task has high or low diagnostic potential and whether selecting a specific task in a specific simulation reflects sensitivity to diagnostic potential. *Adaptivity* describes whether selecting the task was considered to be an adaptive choice for the corresponding student case profile (only for tasks from the second phase of the interview). Sections Coding of diagnostic potential (sensitivity) and Coding

of adaptivity of task selection for each student case profile provide a more detailed explanation of how sensitivity and adaptivity were coded.

We analyzed the data on the level of individual task selections, taking its nested structure into account. To achieve this, we estimated generalized linear mixed models (Bates et al., 2014) to predict the probability that a specific participant would select a specific task in a specific interview. The interview position (first *vs*. second interview) and its interaction with the other fixed factors were included in all models.

To investigate participants' sensitivity to diagnostic task potential (RQ1.1, RQ1.2), we included the tasks' *diagnostic potential* (low *vs*. high) as a fixed factor. The generalized linear mixed model (GLMM) estimators for this effect describe the logarithmized odds ratio for selecting a high-potential task *vs*. selecting a low-potential task.

To investigate the relationship between sensitivity and participants' professional knowledge (RQ1.2), professional knowledge scores and their interaction effect with the *diagnostic potential* factor were additionally included as fixed effects. The GLMM estimate of the interaction term between *diagnostic potential* and professional knowledge scores describes how much the logarithmized odds ratio to select a high-potential task (compared to a low potential-task) changes if professional knowledge scores increase by one standard deviation.

For the questions about participants' adaptive use of diagnostic potential (RQ2.1, RQ2.2), only task selections from the second phase were considered. We assumed that adaptivity builds on insights that could have been generated in prior observations during the first phase of the interview. In the corresponding GLMM, the *adaptivity* of task selection (non-adaptive *vs*. adaptive) was included as a fixed factor. The estimate corresponding to this factor describes the logarithmized odds ratio for making an adaptive *vs*. non-adaptive task selection.

The relationship with participants' professional knowledge (RQ2.2) was again analyzed by including the professional knowledge scores and their interaction effect with the *adaptivity*.

Regarding the models' random effects structure, random intercepts were included to account for differences between individual participants, the four student case profiles, and the different tasks. To investigate whether sensitivity to diagnostic potential or adaptive use of diagnostic potential varied systematically between persons (RQ1.1, RQ2.1), random slopes of *diagnostic potential* resp. *adaptivity* varying over individual participants were included. If the model with this random slope showed a better fit to the data than a model without it, this indicates that participants do indeed systematically vary in their preference for high-potential tasks over low-potential tasks (resp. in the adaptive use of task potential). Since participants' sensitivity and adaptivity might depend on the student case profile, we also analyzed random slopes of *diagnostic potential* resp. *adaptivity* varying over student case profiles. Random slopes were removed from the models before the main analysis if they did not contribute significantly to model fit. Random effects with zero variance estimators were also removed from the models.

Model comparisons were performed with chi-square difference tests. Fixed effects were analyzed with Type-III Wald chi-square tests. Statistical analyses were computed using R and the package lme4 (Bates et al., 2014).[1]

Instruments

Coding of Diagnostic Potential (Sensitivity)

To assess the participants' sensitivity to the diagnostic potential of tasks, each task was coded as having low or high potential. The coding was performed by experienced mathematics education researchers, based on research on students' understanding of decimal fractions, (e.g. Steinle, 2004). Within task clusters, the coding of potential also relied on the comparison of the tasks included in this specific cluster, with distinctions as to whether there was another task with a higher suitability for diagnosing competences or misconceptions in the field of decimal fractions. Two independent coders rated all the tasks. Discrepancies were resolved through discussion among the two coders and a third member of the research group. Four of the ten initial tasks and 16 of the 35 subsequent tasks were coded with high potential. The order of tasks was not linked to task potential.

Coding of Adaptivity of Task Selection for Each Student Case Profile

All tasks of the second phase of the interview were additionally coded as adaptive or non-adaptive independently for each student case profile. The concept of adaptivity takes into account whether a single task is appropriate for delivering further information. Consequently, the coding of adaptivity varies according to the different student case profiles. The coding of adaptivity was independent of the coding of sensitivity, but the method for coding adaptivity considered the indicated diagnostic possibility of a single task. This general suitability was then valued as to whether it could generate additional evidence based on what could have been observed in the screening tasks.

Adaptivity coding was performed by the same two coders independently from one another and separately for each student case profile. As previously mentioned, discrepancies were addressed through discussions among the two coders and a third member of the research group. From the 35 subsequent tasks, 16 were coded as adaptive for Student Case Profile 1 and 4, 12 for Student Case Profile 2, and 20 for Student Case Profile 3. The order of tasks was not linked to adaptivity coding.

Professional Knowledge

Participants' professional knowledge was measured following the categorization of Schulman (1987). Twelve items were used for CK. The scale assessed mathematical knowledge of decimal fractions on a level that required substantial reflection on school mathematics. For example, participants had to justify (without using the usual calculation rules for decimal fractions), that $0.3 \times 0.4 = 0.12$ (**Supplementary Figure S1** in the Supplementary Material). PCK

[1]Additional information about the instruments, the dataset, and the analysis of the data are available from the authors on request.

TABLE 1 | Item statistics.

		CK	PCK	PK
Whole sample (N = 357)	# Of items, whole test	24	16	11
	EAP reliability	0.60	0.58	0.55
	Item-parameters			
	# Of items, this study	12	8	11
	M	1.07	0.66	0.90
	SD	1.06	0.88	0.83
Scaling sample (N = 292)	Person-parameters			
	M	0.30	0.14	−0.03
	SD	1.05	1.28	1.12
Sample of current study (N = 65)	Person-parameters			
	M	0.98	0.88	0.08
	SD	1.02	0.91	0.87

Item-parameters are only given for items included in the current study.

TABLE 2 | Descriptive statistics for interviews: Mean values (M) and standard deviation (SD) of interview duration [in minutes] and number of selected tasks for each interview and student case profile.

Student case profiles	First interview		Second interview	
	Duration	Number of tasks	Duration	Number of tasks
	M (SD)	M (SD)	M (SD)	M (SD)
Profile 1	25.3 (7.3)	12.8 (7.1)	24.2 (5.5)	17.6 (5.3)
Profile 2	27.1 (5.6)	15.7 (5.9)	24.3 (6.3)	17.7 (8.8)
Profile 3	26.5 (6.7)	13.4 (5.3)	20.3 (8.9)	16.0 (7.9)
Profile 4	27.1 (5.0)	12.8 (4.0)	26.1 (5.7)	13.8 (6.3)
Total	26.5 (6.1)	13.8 (5.7)	23.6 (6.9)	16.3 (7.1)

Maximum duration of interview 30 min, number of available tasks for selection 45.

was measured with eight items, focusing on the teaching and learning of decimal fractions. For example, participants were asked to describe a typical incorrect solution strategy for the division problem 4.8 : 2.2 = … (**Supplementary Figure S2** in the supplementary material). For the CK and PCK items, single choice, multiple choice, and open-ended items were used. The scale for PK was adopted from the KiL project (Kleickmann et al., 2014). As this study focuses on diagnostic competences, only the items covering diagnostic-related knowledge were used. This amounts to 11 items pertaining to knowledge about assessment from a psychological point of view, for example, general judgment errors. All PK items had a multiple choice format.

Data from all the tests were made available for the participants of our study and an additional scaling sample of 292 pre-service mathematics high school teachers studying at the same university. For CK and PCK, the scaling sample covered a larger pool of items (24 CK items and 16 PCK items in total) in a multi-matrix design with four booklets. Individual knowledge scores[2] for each of the three components, as well as scaled characteristics, were calculated for both samples together (**Table 1**, for detailed information see **Supplementary Table S1.1** in the supplementary material) using the one-dimensional one-parameter logistic Rasch model (Rasch, 1960); person-parameters are presented for each sample separately.

RESULTS

Descriptive Findings

On average, the participants' first interview had a duration of 26.5 min (*SD* = 6.1), and the second interview ended a bit earlier, on average, after 23.6 min (*SD* = 6.9). The duration of the interviews did not differ remarkably across the different student case profiles. Participants, on average, selected 13.8 tasks for their first interview (*SD* = 5.7) and 16.3 tasks (*SD* = 7.1) for the second interview, with only small descriptive differences among the student case profiles. Even though the total duration of a single interview decreased from the first to the second interview, the number of selected tasks increased (**Table 2**).

An analysis of task selection across the different interview phases yielded almost no differences in the section of the initial tasks (**Table 3**). On average, participants selected 6.2 initial tasks for their first interview (*SD* = 1.5) and 6.5 for their second interview (*SD* = 1.6). From the 35 subsequent tasks, participants selected 7.6 tasks on average (*SD* = 5.3) for the first interview and 9.8 tasks (*SD* = 6.6) for the second interview. From first informal observations, we noticed that task selection in each task cluster seemed to be influenced by task order: participants often chose the first task in a task cluster first, followed by one or more subsequent tasks in the same task cluster.

Sensitive Task Selection

To investigate participants' sensitivity to the diagnostic potential of tasks, we estimated GLMMs to predict selection (*vs.* non-selection) of a task based on its *diagnostic potential* (low *vs.* high), the *interview position* (first *vs.* second interview), and their interaction as fixed factors. In the initial model, we included random intercepts for each participant, each task, and each student case profile, as well as random slopes of the factor *diagnostic potential* varying over participants and student case profiles (Model 1, **Table 4**). Removing the random slope and intercept over student case profiles (Model 2) did not significantly affect model fit ($\chi^2(3) = 1.61$; $p = .658$). However, removing the random slope of *diagnostic potential* varying over participants (Model 3) would have significantly reduced model fit ($\chi^2(2) = 13.32$; $p = .001$), indicating that participants systematically varied in their tendency to prefer high-potential over low-potential tasks.

Sensitivity to Diagnostic Potential (RQ1.1)

On average, participants selected 7.0 high-potential tasks (*SD* = 3.4) and 8.0 low-potential tasks (*SD* = 3.9). To study the overall sensitivity to diagnostic potential in the interviews, we analyzed the fixed effects in Model 2.

No significant effects of *diagnostic potential* ($\chi^2(1) = 0.04$; $p = .851$) occurred, indicating that the predicted probabilities for selecting a low- or a high-potential task did not

[2]The IRT knowledge scores can be interpreted in the following way: a person with knowledge score θ will, according to the Rasch model, solve an item with difficulty parameter δ with a probability of $p = 1/(1+\exp(\delta-\theta))$.

TABLE 3 | Descriptive statistics for selection of tasks: Mean values (M) and standard deviation (SD) of number of selected tasks, differed by initial and subsequent tasks and student case profiles.

Student case profiles	Initial tasks		Subsequent tasks	
	First interview	Second interview	First interview	Second interview
	M (SD)	M (SD)	M (SD)	M (SD)
Profile 1	6.5 (1.9)	6.8 (0.9)	6.3 (6.3)	10.8 (5.2)
Profile 2	6.4 (1.3)	6.6 (1.5)	9.3 (5.5)	11.1 (8.3)
Profile 3	5.8 (1.5)	6.2 (1.9)	7.6 (5.1)	9.8 (7.2)
Profile 4	6.1 (1.4)	6.3 (1.9)	6.7 (3.8)	7.4 (5.9)
Total	6.2 (1.5)	6.5 (1.6)	7.6 (5.3)	9.8 (6.6)

Total number of available initial tasks 10, total number of available subsequent tasks 35.

differ significantly. The effect of *interview position* was significant ($\chi^2(1) = 11.10; p < .001$) indicating that during the second interview, the participants chose significantly more tasks than in the first interview. Finally, the interaction effect between *diagnostic potential* and *interview position* was not significant ($\chi^2(1) = 0.97; p = .326$), indicating that the selection of more tasks in the second interview was independent of the tasks' diagnostic potential.

Professional Knowledge and Sensitivity to Diagnostic Potential (RQ1.2)

To investigate the role of professional knowledge in sensitivity to diagnostic potential, we included participants' scores for each professional knowledge component (CK, PCK, PK) as well as interaction effects with *diagnostic potential* separately in Models 4 to 6.

For CK (Model 4), there was a significant main effect of the professional knowledge score ($\chi^2(1) = 5.18; p = .023$) and a significant interaction with *diagnostic potential* ($\chi^2(1) = 4.62; p = .032$). Trend analyses (**Figure 3**) showed that the selection of low-potential tasks was significantly negatively related to CK [$B = -0.24$, $CI_{95\%}$ (−0.45,−0.03)], while the selection of high-potential tasks was not significantly related to CK [$B = -0.06$, $CI_{95\%}$ (−0.28,0.16)]. The difference between the two (logistic) regression slopes was significant ($B = 0.18$, $p = .030$), indicating that when CK scores were one standard deviation higher, they were accompanied by approximately 19.3% higher odds of selecting a high-potential task, as opposed to a low-potential task (**Figure 3**).

For PCK (Model 5) and PK (Model 6), neither the main effects of the knowledge scores (PCK: $\chi^2(1) = 0.25$; $p = .620$; PK: $\chi^2(1) = 0.00$; $p = .956$) nor the corresponding interactions with diagnostic potential (PCK: $\chi^2(1) = 0.53$; $p = .466$; PK: $\chi^2(1) = 0.49$; $p = .486$) were significant.

Adaptive Task Selection

In investigating the adaptive use of diagnostic task potential, only second-phase task selection was considered. In the second phase of the interview, participants had the opportunity to generate further evidence based on what they had observed during the first phase of the interview. The estimated GLMMs predict the selection (*vs.* non-selection) of a task based on its *adaptivity* to the related student case profile, the *interview position* (first *vs.* second interview), and their interaction as fixed factors. In the initial model, we included random intercepts for each participant, each task, and each student case profile, as well as random slopes of the factor *adaptivity* varying over participants and student case profiles (Model 1, **Table 5**). Removing the random slope and intercept over student case profiles (Model 2) did not significantly affect model fit ($\chi^2(3) = 2.36; p = .501$), as well as removing the random slope and intercept over participants (Model 3), indicating that participants did not systematically vary in their tendency to select tasks adaptively ($\chi^2(2) = 1.77; p = .413$).

Adaptive Use of Diagnostic Task Potential (RQ2.1)

To analyze the fixed effects in terms of the adaptive use of diagnostic task potential, Model 3 was investigated.

The effect of *adaptivity* was not significant ($\chi^2(1) = 0.06; p = .811$), indicating that participants did not systematically take the characteristics of the student case profile into account when selecting tasks, at least not regarding adaptivity as operationalized above. The effect of the *interview position* was significant ($\chi^2(1) = 15.77; p < .001$), whereas the interaction of *adaptivity* and *interview position* had no effect ($\chi^2(1) = 0.00; p = .953$). Together, these results indicate that task selection changed from the first to the second interview, without being more adaptive to the student case profile.

Professional Knowledge and Adaptive Use of Diagnostic Task Potential (RQ2.2)

To investigate the role of professional knowledge in the adaptive use of diagnostic potential, we included participants' scores separately for each professional knowledge component (CK, Model 4; PCK, Model 5; PK, Model 6) as well as interaction effects with *adaptivity*.

For none of the professional knowledge components significant effects occurred. Neither the main effects of the knowledge scores (CK: $\chi^2(1) = 2.06; p = .151$; PCK: $\chi^2(1) = 0.27; p = .602$; PK: $\chi^2(1) = 0.28; p = .599$) nor the corresponding interactions with *diagnostic potential* (CK: $\chi^2(1) = 0.70; p = .404$; PCK: $\chi^2(1) = 0.12; p = .728$; PK: $\chi^2(1) = 0.30; p = .585$) were significant.

DISCUSSION

This article focuses on teachers' task selection during diagnostic one-to-one simulations and investigates the study participants' sensitivity to and adaptive use of the diagnostic potential of tasks

TABLE 4 | Different GLMM for analyzing the sensitivity to the diagnostic task potential.

Fixed Effects	**Model 1**			**Model 2**			**Model 3**			**Model 4**			**Model 5**			**Model 6**		
	B	χ^2	***p***	***B***	χ^2	***p***	***B***	χ^2	***p***	***B***	χ^2	***p***	***B***	χ^2	***p***	***B***	χ^2	***p***
Diagnostic potential	0.08	0.04	.842	0.07	0.04	.851	0.07	0.04	.850	–0.11	0.08	.777	0.01	0.00	.985	0.06	0.03	.864
Interview position	**0.31**	**11.81**	**<.001**	**0.29**	**11.10**	**<.001**	**0.29**	**10.97**	**<.001**	**0.29**	**11.10**	**<.001**	**0.29**	**11.10**	**<.001**	**0.29**	**11.10**	**<.001**
Diagnostic potential x interview position	0.12	0.84	.358	0.13	0.97	.326	0.13	0.95	.330	0.13	0.96	.326	0.13	0.97	.326	0.13	0.97	.326
CK										**–0.24**	**5.18**	**.023**						
PCK													–0.06	0.25	.620			
PK																–0.01	0.00	.956
Diagnostic potential x CK										**0.18**	**4.62**	**.032**						
Diagnostic potential x PCK													0.07	0.53	.466			
Diagnostic potential x PK																0.07	0.49	.486
Random effect variances	*Var*	*SD*		*Var*	*SD*		*Var*	*SD*		*Var*	*SD*		*Var*	*SD*		*Var*	*SD*	
Task: Intercept	1.43	1.19		1.43	1.19		1.39	1.18		1.42	1.19		1.43	1.19		1.43	1.19	
Participant: Intercept	0.67	0.82		0.68	0.83		0.62	0.79		0.62	0.79		0.68	0.82		0.68	0.83	
Participant: Slope of diagnostic potential	0.22	0.47		0.22	0.47					0.20	0.44		0.22	0.47		0.22	0.47	
Student case profile: Intercept	0.01	0.11																
Student case profile: Slope of diagnostic potential	0.00	0.07																
R^2																		
Marginal	0.01			0.01			0.01			0.01			0.01			0.01		
Conditional	0.35			0.35			0.32			0.35			0.35			0.35		
AIC	5,996.8			5,992.4			6,001.7			5,989.4			5,995.8			5,995.9		
BIC	6,070.2			6,045.8			6,041.8			6,056.1			6,062.6			6,062.6		
Deviance	5,974.8			5,976.4			5,989.7			5,969.4			5,975.8			5,975.9		

Significant effects (p < .05) were highlighted in bold; df of χ^2-tests equal 1.

when diagnosing students' mathematical understanding. Live simulations of diagnostic one-to-one interviews were used to ensure authentic but comparable conditions. The importance of diagnostic competences (Herppich et al., 2018; Heitzmann et al., 2019; Loibl et al., 2020) and knowledge about the diagnostic potential of tasks have been put forward in the literature (Baumert & Kunter, 2013). The need for approaches to studying diagnostic processes under controlled but authentic settings has been raised as a desiderate (Grossman & McDonald, 2008). We propose and investigate an innovative perspective on the diagnostic process, focusing on the selection of diagnostic tasks. Differentiating between sensitivity to and the adaptive use of diagnostic task potential takes into account that a task can have more than just high or low diagnostic potential; that is, a task's potential may be more or less useful depending on what is already known about a specific student's mathematical understanding in a diagnostic process. In our study, we were able to investigate pre-service teachers' sensitivity to and adaptive use of diagnostic task potential during authentic (though, of course, not real) diagnostic situations in a controlled setting. Even though the importance of diagnostic competences is undisputed (Behrmann & Souvignier, 2013) the constructs of sensitivity to and the adaptive use of diagnostic task potential during the diagnostic process have not been described in detail in the literature, and evidence from authentic diagnostic processes under controlled conditions is scarce. Our study provides results for the measurement of these characteristics, pre-service teachers' sensitivity to and adaptive use of diagnostic task potential and the relationship with professional knowledge.

Measurement of the Proposed Process Characteristics

Knowledge and use of diagnostic task potential have been underlined as important aspects of teachers' professional competences in the past (Baumert & Kunter, 2013), and their operationalization and measurement have been discussed repeatedly. For example, Herppich et al. (2018) call for a wider spectrum of criteria to assess diagnostic competence, including process-based measures. In response to this, the present study examined pre-service teachers' sensitivity to and adaptive use of the diagnostic potential of tasks when diagnosing students' mathematical understanding. Regarding participants' sensitivity, our results partially met our expectations, as implied by prior research. However, there was no main effect of sensitivity, indicating that participants generally did not have a higher probability of selecting tasks with high diagnostic potential. The results reveal a systematic inter-individual variation between the participants in their tendency to prefer

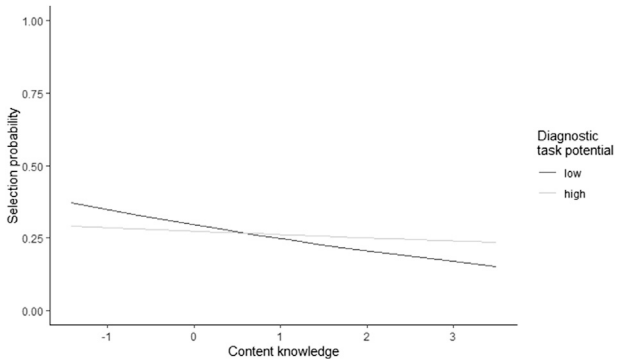


FIGURE 3 | The relationship between CK and the selection probability of high-*vs*. low-potential tasks.

high-potential over low-potential tasks. The observed systematic variation indicates that the construct of sensitivity and its operationalization in our study might indeed reflect meaningful characteristics of the diagnostic process, which reflect the participants' diagnostic competences.

The picture is different for the adaptive use of diagnostic potential, which we had already expected to be a more complex characteristic. First, the adaptive use of diagnostic potential requires complex cognitive processes, including the interpretation and integration of information from prior student answers, which may differ inter-individually. Second, the measurement of this construct is intricate, as participants' prior information about the student cannot be accessed explicitly, but a proxy measure had to be used, assuming that participants had collected at least the basic information that could have been observed during the first phase of the interview based on the first three task clusters. In our study, we could not identify a significant direct effect of adaptivity or inter-individual variation in the adaptive use of task potential. This can be for different reasons. First, it might be that the demand for using tasks adaptively was just beyond what our participants could achieve, given that at the time of study, they were just mid-way through their university studies and had limited practical experience. In other words, the required cognitive processes might have been too complex for the participants. Second, it might be that our operationalization of the construct as a proxy measure was too coarse to capture the situation-specific adaptivity of teachers' task selections. In this case, it remains an open question if and how more valid, but still efficient measures of adaptivity might be developed, for example, by explicitly asking participants about intermediate diagnoses or with more effort, by analyzing prior interactions in the interview qualitatively. Moreover, adaptive task selection might be influenced much more strongly by situation-specific factors than by individual dispositions, making finding systematic inter-individual variance impossible. In this case, the construct (possibly with improved operationalization) might even have additional predictive value for the final diagnosis in the sense of a situation-specific skill (Blömeke et al., 2015) beyond stable person characteristics. Based on our data, we cannot provide evidence for or against these assumptions. A good way to study this issue would be to investigate whether and how adaptivity (as measured in our study or with alternative operationalization) goes along with better, for example more accurate diagnoses.

Pre-Service Teachers' Sensitivity to and Adaptive Use of Diagnostic Task Potential

Beyond establishing and measuring the construct, our study yields first evidence as to what degree pre-service teachers are indeed sensitive to the diagnostic potential of tasks. The results show that the odds of choosing high-potential tasks (over low-potential tasks) did not differ significantly. This indicates that on average, our participants did not prefer high-potential tasks, and that selecting tasks according to their diagnostic potential was not straightforward for our sample. Even though some participants seemed to systematically prefer high-potential tasks (significant systematic inter-individual differences), others did not. Thus, at least the latter participants, but also other pre-service teachers, might require further support in attending to task potential and developing their sensitivity. Prior studies have shown that dedicated training in the PCK component of professional knowledge can improve teachers' ability to estimate task difficulty (Ostermann et al., 2018). Our study cannot provide evidence as to whether this is similar for sensitivity to tasks' diagnostic potential of tasks. However, it can provide first insights into the role of all three components of professional knowledge (see below).

As for the adaptive use of diagnostic task potential, we did not find a significant effect on task selection or significant systematic inter-individual variation between participants. Thus, it appears that participants were generally not significantly more likely to make adaptive *vs*. non-adaptive task selections, implying that participants had difficulty with the adaptive use of diagnostic task potential. Since being sensitive to diagnostic task potential is a prerequisite to the adaptive use of task potential in our process model, it seems plausible that support would have to address strategies and knowledge pertaining to both to foster the ability to detect the diagnostic potential of a task and judge it against what is already known about a student.

Professional Knowledge

The measures used for professional knowledge only showed significant relationships with sensitivity to diagnostic task potential, but no relationship with the adaptive use of diagnostic task potential. However, as no significant systematic inter-individual variation could be observed for the adaptive use of diagnostic task potential, this came as no surprise given that professional knowledge is a person characteristic. Other factors, such as situation-specific motivational states (e.g., Herppich et al., 2018), might moderate these relationships and substantially reduce the bivariate correlations. Regarding sensitivity to diagnostic task potential, we primarily expected a relationship with participants' PCK. This assumption was based on the fact that knowledge about the diagnostic potential of tasks is often discussed as a facet of PCK (Baumert & Kunter, 2013), and prior findings on the relationship of teachers' PCK and their diagnostic competence (Ostermann et al., 2018). Moreover, although

TABLE 5 | Different GLMM for analyzing the adaptive use of diagnostic task potential.

Fixed Effects	**Model 1**			**Model 2**			**Model 3**			**Model 4**			**Model 5**			**Model 6**		
	B	χ^2	***p***	***B***	χ^2	***p***	***B***	χ^2	***p***	***B***	χ^2	***p***	***B***	χ^2	***p***	***B***	χ^2	***p***
Adaptivity	0.00	0.00	.999	-0.02	0.00	.946	-0.05	0.06	.811	-0.11	0.22	.643	-0.03	0.02	.902	-0.05	0.05	.820
Interview position	**0.43**	**16.20**	**<.001**	**0.41**	**15.90**	**<.001**	**0.41**	**15.77**	**<.001**	**0.41**	**15.90**	**<.001**	**0.41**	**15.78**	**<.001**	**0.41**	**15.76**	**<.001**
Adaptivity x interview position	-0.01	0.00	.957	0.00	0.00	.987	0.01	0.00	.953	0.01	0.00	.964	0.01	0.00	.949	0.01	0.00	.952
CK										-0.22	2.06	.151						
PCK													-0.09	0.27	.602			
PK																-0.10	0.28	.599
Adaptivity x CK										0.06	0.70	.404						
Adaptivity x PCK													-0.03	0.12	.728			
Adaptivity x PK																-0.05	0.30	.585
Random effect variances	*Var*	*SD*		*Var*	*SD*		*Var*	*SD*		*Var*	*SD*		*Var*	*SD*		*Var*	*SD*	
Task: Intercept	0.40	0.63		0.40	0.63		0.40	0.63		0.40	0.63		0.40	0.63		0.40	0.63	
Participant: Intercept	1.53	1.24		1.54	1.24		1.42	1.19		1.37	1.17		1.41	1.19		1.42	1.19	
Participant: Slope of adaptivity	0.09	0.30		0.08	0.29													
Student case profile: Intercept	0.01	0.10																
Student case profile: Slope of adaptivity	0.01	0.12																
R^2																		
Marginal	0.01			0.01			0.01			0.01			0.01			0.01		
Conditional	0.31			0.31			0.26			0.26			0.26			0.26		
AIC	4,474.9			4,471.3			4,469.1			4,470.8			4,472.6			4,472.3		
BIC	4,545.6			4,522.7			4,507.6			4,522.1			4,524.0			4,523.7		
Deviance	4,452.9			4,455.3			4,457.1			4,454.8			4,456.6			4,456.3		

Significant effects (p < .05) were highlighted in bold; df of χ^2-tests equal 1.

diagnostic potential is intrinsically a very content-related task characteristic, it is not only based on a task's mathematical characteristics, but also on the strategies that can be applied to solve the task. However, in our study, only participants' CK correlated with sensitivity to diagnostic task potential. This parallels the findings of Van den Kieboom et al. (2013) who noted that pre-service teachers' CK was related to their questioning behavior in diagnostic interviews. It seems that for the participants in our sample, a preference for high-potential tasks relied more on their sound understanding of decimal fractions than on their knowledge about student cognition and instruction. In this context, it is an interesting result that higher CK scores did not significantly go along with higher odds of selecting high-potential tasks, but rather with lower odds of selecting low-potential tasks. CK thus appears to enable students to identify and discard low-potential tasks, but it is not sufficient to facilitate the identification and implementation of high-potential tasks. The relatively high impact of CK appears plausible, as superficial strategies, which lower the diagnostic potential of tasks, might be detected by two different methods: 1) recognition of well-known superficial strategies that students apply when dealing with decimals, which would be part of PCK, or 2) a mathematical analysis of as many strategies as possible to solve the task, which would theoretically rely primarily on CK. Our results indicate that our participants applied the second CK-based strategy to a larger extent. However, efficient selection of diagnostic tasks in everyday practice would plausibly be easier to achieve with the first strategy, as mathematical analysis can be expected to be more demanding and time consuming. Since prior studies have not identified a spontaneous transfer of learned CK to PCK tasks (Tröbst et al., 2019) changing to more efficient strategies might still rely on learning PCK. Training studies, but also analyzing diagnostic processes as used by practicing teachers, might provide first insights as to whether stronger or more enriched PCK might lead to different strategies. In this regard, specifically the distinction between personal PCK, (i.e. acquired knowledge related to the diagnostic situation) and enacted PCK, (i.e. knowledge actually used in diagnostic situations), following Carlson and Daehler (2019) refined consensus model, might be of high relevance. In particular, it is likely that our participants acquired the necessary PCK for the simulated diagnostic interview, but did not enact the required PCK for the situation, thus failing to put their theoretical knowledge into action.

Student Case Profiles

Differences between the four student case profiles were controlled in this study, mainly to avoid potential distortions of the results. However, it is interesting that the participants' tendency to select high-potential tasks and make adaptive task selections did not vary systematically across the four student case profiles. For measurement purposes, this indicates that the four student case profiles can be used mostly interchangeably, as they do not show substantially different levels of difficulty regarding the two characteristics of the diagnostic process.

First *vs*. Second Interview

The results were mostly comparable for the first and the second interview, as far as participants' tendency to select high-potential tasks and make adaptive task selections is concerned. In particular, we did not observe any short-term learning effects regarding either of the process measures. Beyond this, however, it seems that pre-service teacher' task selection behavior and diagnostic processes did change from the first to the second interview. In particular, participants selected more (high- and low-potential and adaptive and non-adaptive) tasks in the second interview, but spent less time on these tasks, on average. Based on the data analyzed in this contribution, it has to remain an open question whether the reason for this is more efficient task presentation and diagnostic interpretation of students' responses or whether it reflects more superficial work during the second simulation. Again, studying the relationship between pre-service teachers' task selection, their diagnostic interpretations drawn from the task during the interview, and the accuracy of their final evaluation of the student's understanding could provide a path to obtaining deeper insights into the mechanisms at work behind these differences. The observed differences themselves, however, point to the fact that investigating the effects of repeated engagement in simulations in pre-service teacher education should be carefully investigated in the future and may add interesting results regarding learning effects beyond single encounters with such situations.

The findings of this study show that without further support, pre-service teachers do not select tasks sensitively regarding their diagnostic potential or even adopt task selection in accordance with information that has already been gathered about the student's understanding. From an informal perspective, it seems that the participants based their task selection more on aspects of task presentation (in particular, their order of presentation), than on task characteristics connected to diagnostic potential. The significant effect of CK points out that supporting pre-service teachers' professional knowledge could be promising. The fact that the participants in our sample had already encountered all the necessary CK, PCK, and PK content for the simulations in lectures and small group tutorials points to the fact that simply acquiring the relevant knowledge (pPCK) might not be sufficient for enacting (ePCK) the knowledge in a diagnostic situation. Contrary to CK, which pre-service teachers have already encountered to some extent in their own school careers, the application of PCK and PK in particular might rely on sufficient learning opportunities in authentic situations, such as simulations (used as learning environments) and practical studies. Thus, the inclusion of authentic applications of acquired knowledge in university studies is of central interest, as well as how diagnostic processes, in particular, including task selection, can be supported in such settings. Investigating the effects of prompts (Berthold et al., 2007) and reflection phases (Mamede et al., 2012) could also provide differentiated insights about the role of professional knowledge. Assuming that the selection of tasks is a key part of the diagnostic process, fostering pre-service teachers' sensitivity to and adaptive use of diagnostic task potential when selecting tasks (Moyer-Packenham & Milewicz, 2002) seems to be very promising.

LIMITATIONS

Of course, our study suffers from a number of limitations. Most importantly, the operationalization of the concept of adaptivity in the present study is quite limited. The construct of an adaptive use of diagnostic task potential builds on the insights that a participant gains at a specific point in a diagnostic interview. Thus, direct operationalization would need to build on individual information reconstructed by the participant, which cannot be systematically controlled. As described above, alternative approaches to measure adaptivity in diagnostic task selection, but also further analyses of the current operationalization might be promising for addressing the open issues connected to this construct.

Moreover, we introduce new process characteristics and investigate them in a very specific setting, spanning over four student case profiles, which, despite their differences, are all based on the same pool of diagnostic tasks from the content area decimal fractions. It thus remains an open question to what extent our results would transfer not only to different mathematical content, but also to different diagnostic situations and different populations of pre- and in-service teachers (Karst et al., 2017).

The chosen setting itself can be seen as valid for teachers, since diagnosing individual students' understanding of a specific concept using mathematical tasks is part of teachers' everyday practice. However, it must be taken into account that one-to-one interviews lasting about 30 min each are not feasible as everyday practice in many schools. On the other hand, this choice allowed us to generate a controlled yet sufficiently authentic setting to investigate our questions and gather a sufficient amount of data, which would not have been possible in less time. In particular, Grossman et al. (2009) propose the use of such approximations of practice as learning opportunities in pre-service teacher education, and Shavelson (2012) argues for their use as assessment tools. In this sense, the results are of interest for the development of such practice approximations, even though they are not broadly part of everyday teacher practice.

Due to the sample size of 65 participants, the insignificant or almost significant effects could be explained by restricted statistical power in our study. Investigating this approach based on a larger sample size could be promising. In addition, a comparison of pre- and in-service teachers' task selection could lead to clearer contrasts between both groups and provide auspicious insights into how 1) pre-*vs*. in-service teachers shape their diagnostic processes and select tasks, and how 2) pre- and in-service teachers' pPCK and ePCK are related to their sensitivity to and adaptive use of diagnostic task potential. These insights could be the basis for future professional development. Moreover, a replication with a larger, more representative sample, possibly also from different universities or countries, would help to support our findings about the absolute level of participants' sensitivity. However, because inter-individual differences in performance are not systematically linked to a specific sample's performance level, stronger generalizability can be assumed for these findings.

Finally, since only professional knowledge was considered as a participant prerequisite, investigating other trait

prerequisites, such as interest or state variables like motivation, authenticity, or cognitive demand (Codreanu et al., 2020), would allow for more differentiated insights, for example, by considering the interplay of knowledge and motivation, and focusing on moderating effects that may obscure the effects of participants' professional knowledge.

CONCLUSION

This study presents a role-play based live simulation that was used to assess mathematics pre-service teachers' diagnostic competences in an authentic setting. Participants acted like real teachers, trying to diagnose students' mathematical understanding. Since the students were played live by trained teaching assistants, the participants had a huge scope of action, (e.g. select tasks, interact with the student, pose follow-up questions), while still ensuring the comparability of the experiences within the simulation. The use of this authentic simulation thus enabled the investigation of a close approximation of the participants' natural behavior. By focusing on the participants' selection of tasks during diagnostic interviews, sensitivity to and adaptive use of diagnostic task potential were analyzed as well as the relationship with participants' professional knowledge. Being sensitive to the diagnostic potential of tasks reflects that a task's diagnostic potential is considered to be an important factor during task selection, whereas the construct of the adaptive use of diagnostic task potential additionally reflects the influence of previously collected information about the student's understanding. The results show that sensitivity to diagnostic task potential seems to be related to participants' CK.

This study provides insights into the repeatedly highlighted (Moyer-Packenham & Milewicz, 2002; Maier et al., 2010) yet quite under-investigated role of diagnostic task potential. Differentiating between sensitivity to and adaptive use of diagnostic task potential, this study focuses on the diagnostic process from an innovative perspective. The simulation-based approach in our study facilitated an investigation of the use of diagnostic task potential in task selection during diagnostic processes for the first time in an authentic empirical setting. The findings of this study underline the need for learning environments to foster pre-service teachers' diagnostic competences as well as their underlying professional knowledge. In particular, a major focus should be on enabling them to apply their professional knowledge in appropriate authentic settings in order to develop sensitivity to tasks' diagnostic potential, and make adaptive use of that potential, so that they are able to individually address their prospective students and create custom-tailored, effective diagnostic processes that will be beneficial to both, students and teachers. The findings of this study show that even basic aspects of diagnostic competence, such as the selection of diagnostic tasks, are related to the knowledge pre-service teachers acquire in their university courses. However, at the early stages of this development, it seems that it is not primarily PCK that plays a role in identifying tasks with high diagnostic potential, but CK is required to dismiss tasks with low diagnostic potential. A possible reason could be that students rely more on mathematical analysis of the tasks, rather than their knowledge about possible misconceptions. Accordingly, instructional approaches, like the use of simulations, should ensure that students activate and use their PCK, in addition to CK, to describe and improve their diagnostic actions.

Finally, the established simulation was designed to function as an assessment tool (as used in this study) and as a learning environment to foster pre-service teachers' diagnostic competences. Future intervention studies will provide additional insights into how pre-service teachers can be supported effectively using this simulation to increase their diagnostic and assessment competences.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, upon request without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethics committee of the Faculty of Mathematics, Computer Science and Statistics, LMU Munich.

AUTHOR CONTRIBUTIONS

SK participated in the design of the simulation, creation of the employed scales, collected the data, performed the main analyses, and wrote the manuscript. DS supported during the design of the simulation, the creation of the scales for professional knowledge, data analysis and revised the analysis scripts and the manuscript. MA supported during data collection and revised the manuscript. SU initiated the project and supported the design of the simulation, during data collection, and during data analysis. He revised the analyses and the manuscript.

FUNDING

The research presented in this contribution was funded by a grant of the DFG to Stefan Ufer, Kathleen Stürmer, Christof Wecker, and Matthias Siebeck (Grant numbers UF59/5-1 and UF59/5-2 as part of COSIMA, FOR2385).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2021.604568/full#supplementary-material.

REFERENCES

Aufschnaiter, C. v., Cappell, J., Dübbelde, G., Ennemoser, M., Mayer, J., Stiensmeier-Pelster, J., et al. (2015). Diagnostische Kompetenz. *Theoretische Überlegungen zu einem zentralen Konstrukt der Lehrerbildung*. 61 (5), 738–758.

Ball, D. L., Thames, M. H., and Phelps, G. (2008). Content knowledge for teaching: what makes it special? *J. Teach. Educ.* 59, 389–407. doi:10.1177/0022487108324554

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *J. Stat. Soft.* 67 1–103. doi:10.18637/jss.v067.i01

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *Am. Educ. Res. J.* 47 (1), 133–180. doi:10.3102/0002831209345157

Baumert, J., and Kunter, M. (2013). "The COACTIV model of teachers' professional competence," in *Cognitive activation in the mathematics classroom and professional competence of teachers: results from the COACTIV project*. Editors M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, and M. Neubrand (New York: Springer), 25–48.

Behrmann, L., and Souvignier, E. (2013). The relation between teachers' diagnostic sensitivity, their instructional activities, and their students' achievement gains in reading. *Z. für Paedagogische Psychol.* 27 (4), 283–293. doi:10.1024/1010-0652/a000112

Berthold, K., Nückles, M., Renkl, A., and Instruction (2007). Do learning protocols support learning strategies and outcomes? The role of cognitive and metacognitive prompts. *Learn. InStruct.* 17 (5), 564–577. doi:10.1016/j.learninstruc.2007.09.007

Black, P., and Wiliam, D. (2009). Developing the theory of formative assessment. *Educ. Assess. Eval. Account.* 21 (1), 5. doi:10.1007/s11092-008-9068-5

Blömeke, S., Gustafsson, J.-E., and Shavelson, R. J. (2015). Beyond dichotomies: viewing competence as a continuum. *Z. für Psychol.* 223, 3–13. doi:10.1027/2151-2604/a000194

Blömeke, S., Kaiser, G., and Lehmann, R. (2011). "Messung professioneller Kompetenz angehender Lehrkräfte: "Mathematics Teaching in the 21st Century" und die IEA-Studie TEDS-M," in *Empirische Fundierung in den Fachdidaktiken*. Editors H. Bayrhuber, U. Harms, B. Muszynski, B. Ralle, M. Rothgangel, L.-H. Schön, et al. (Münster, Germany: Waxmann), 9–26.

Bromme, R. (1981). *Das Denken von Lehrern bei der Unterrichtsvorbereitung: eine empirische Untersuchung zu kognitiven Prozessen von Mathematiklehrern*. Weinheim, Germany: Beltz.

Carlson, J., and Daehler, K. (2019). "The refined consensus model of pedagogical content knowledge in science education," in *Repositioning pedagogical content knwoledge in Teachers´ knwoledge for teaching science*. Editors A. Hume, R. Cooper, and A. Borowski (Singapore: Springer).

Codreanu, E., Sommerhoff, D., Huber, S., Ufer, S., and Seidel, T. (2020). Between authenticity and cognitive demand: finding a balance in designing a video-based simulation in the context of mathematics teacher education. *Teach. Teach. Educ.* 95, 103146. doi:10.1016/j.tate.2020.103146

Förtsch, C., Sommerhoff, D., Fischer, F., Fischer, M., Girwidz, R., Obersteiner, A., et al. (2018). Systematizing professional knowledge of medical doctors and teachers: development of an interdisciplinary framework in the context of diagnostic competences. *Educ. Sci.* 8 (4), 207. doi:10.3390/educsci8040207

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., and Williamson, P. (2009). Teaching practice: a cross-professional perspective. *Teach. Coll. Rec.* 111 (9), 2055–2100.

Grossman, P., and McDonald, M. (2008). Back to the future: directions for research in teaching and teacher education. *Am. Educ. Res. J.* 45 (1), 184–205. doi:10.3102/0002831207312906

Heckmann, K. (2006). *Zum Dezimalbruchverständnis von Schülerinnen und Schülern: theoretische Analyse und empirische Befunde*. Berlin, Germany: Logos.

Heinze, A., and Verschaffel, L. (2009). Flexible and adaptive use of strategies and representations in mathematics education. *ZDM Mathematics Education* 41, 535–540. doi:10.1007/s11858-009-0214-4

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, C., Wecker, M. R., Fischer, S., et al. (2019). Facilitating diagnostic competences in simulations in higher education A framework and a research agenda. *Flr* 7, 1–24. doi:10.14786/flr.v7i4.384

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., et al. (2018). Teachers' assessment competence: integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teach. Teach. Educ.* 76, 181–193. doi:10.1016/j.tate.2017.12.001

Kaiser, J., Praetorius, A.-K., Südkamp, A., and Ufer, S. (2017). "Die enge Verwobenheit von diagnostischem und pädagogischem Handeln als Herausforderung bei der Erfassung diagnostischer Kompetenz," in *Diagnostische Kompetenz von Lehrkräften: theoretische und methodische Weiterentwicklungen*. Editors A. Südkamp and A.-K. Praetorius (Münster, Germany: Waxmann), 114–123.

Karst, K., Klug, J., and Ufer, S. (2017). "Strukturierung diagnostischer Situationen im inner- und außerunterrichtlichen Handeln von Lehrkräften," in *Diagnostische Kompetenz von Lehrkräften: theoretische und methodische Weiterentwicklungen*. Editors A. Südkamp and A.-K. Praetorius (Münster, Germany: Waxmann), 95–113.

Kleickmann, T., Großschedl, J., Harms, U., Heinze, A., Herzog, S., Hohenstein, F., et al. (2014). Professionswissen von Lehramtsstudierenden der mathematisch-naturwissenschaftlichen Fächer-Testentwicklung im Rahmen des Projekts KiL. *Unterrichtswissenschaft* 42 (3), 280–288.

Klug, J., Bruder, S., Kelava, A., Spiel, C., and Schmitz, B. (2013). Diagnostic competence of teachers: a process model that accounts for diagnosing learning behavior tested by means of a case scenario. *Teach. Teach. Educ.* 30, 38–46. doi:10.1016/j.tate.2012.10.004

Koeppen, K., Hartig, J., Klieme, E., and Leutner, D. (2008). Current issues in competence modeling and assessment. *Z. für Psychol./J. Psychol.* 216 (2), 61–73. doi:10.1027/0044-3409.216.2.61

Lin, F.-L., and Rowland, T. (2016). "Pre-service and in-service mathematics teachers' knowledge and professional development," in *The second handbook of research on the psychology of mathematics education: the journey continues*. Editors Á. Gutiérrez, G. C. Leder, and P. Boero (Rotterdam, Netherlands: Sense Publishers), 483–520.

Loibl, K., Leuders, T., and Dörfler, T. (2020). A framework for explaining teachers' diagnostic judgements by cognitive modeling (DiaCoM). *Teach. Teach. Educ.* 91, 103059. doi:10.1016/j.tate.2020.103059

Maier, U., Kleinknecht, M., and Metz, K. (2010). "Ein fächerübergreifendes Kategoriensystem zur Analyse und Konstruktion von Aufgaben," in *Lernaufgaben und Lernmaterialien im kompetenzorientierten Unterricht*. Editors H. Kiper, W. Meints, S. Peters, S. Schlump, and S. Schmit (Stuttgart, Germany: Kohlhammer), 28–43.

Mamede, S., van Gog, T., Moura, A. S., de Faria, R. M., Peixoto, J. M., Rikers, R. M., et al. (2012). Reflection as a strategy to foster medical students' acquisition of diagnostic competence. *Med. Educ.* 46 (5), 464–472. doi:10.1111/j.1365-2923.2012.04217

Marczynski, B., Kaltefleiter, L. J., Siebeck, M., Wecker, C., Stürmer, K., and Ufer, S. (in press). "Diagnosing 6th graders´ understanding of decimal fractions. Fostering mathematics pre-service teachers' diagnostic competences with simulated one-to-one interviews," in *Learning to diagnose with simulations - examples from teacher education and medical education*. Editors F. Fischer and A. Opitz (Heidelberg, Germany: Springer).

Moyer-Packenham, P., and Milewicz, E. (2002). Learning to question: categories of questioning used by preservice teachers during diagnostic mathematics interviews. *J. Math. Teach. Educ.* 5 (4), 293–315. doi:10.1023/A:1021251912775

Neubrand, M., Jordan, A., Krauss, S., Blum, W., and Löwen, K. (2011). "Aufgaben im COACTIV-Projekt: Einblicke in das Potenzial für kognitive Aktivierung im Mathematikunterricht," in *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV*. Editors M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, and M. Neubrand (Waxmann, Germany: Münster u.a.), 115–132.

Ostermann, A., Leuders, T., and Nückles, M. (2018). Improving the judgment of task difficulties: prospective teachers' diagnostic competence in the area of functions and graphs. *J. Math. Teach. Educ.* 21 (6), 579–605. doi:10.1007/s10857-017-9369-z

Padberg, F., and Wartha, S. (2017). *Didaktik der Bruchrechnung*. Berlin, Germany: Springer Spektrum.

Philipp, K. (2018). "Diagnostic competences of mathematics teachers with a view to processes and knowledge resources," in *Diagnostic competence of mathematics teachers: unpacking a complex construct in teacher education and teacher practice*. Editors T. Leuders, J. Leuders, and K. Philipp (New York, NY: Springer), 109–127.

Praetorius, A.-K., Lipowsky, F., and Karst, K. (2012). "Diagnostische Kompetenz von Lehrkräften. Aktueller Forschungsstand, unterrichtspraktische

Umsetzbarkeit und Bedeutung für den Unterricht," in *Differenzierung im mathematisch-naturwissenschaftlichen Unterricht. Implikationen für Theorie und Praxis*. Editors R. Lazarides and A. Ittel (Bad Heilbrunn, Germany: Klinkhardt), 115–146.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Oxford, England: Nielsen & Lydiche.

Schack, E. O., Fisher, M. H., Thomas, J. N., Eisenhardt, S., Tassell, J., and Yoder, M. (2013). Prospective elementary school teachers' professional noticing of children's early numeracy. *J. Math. Teach. Educ.* 16 (5), 379–397. doi:10.1007/s10857-013-9240-9

Schulman, L. S. (1987). Knowledge and teaching: foundations of the new reform. *Harv. Educ. Rev.* 57 (1), 1–23.

Shavelson, R. J. (2012). Assessing business-planning competence using the Collegiate Learning Assessment as a prototype. *Empirical Res Voc Ed Train* 4 (1), 77–90. doi:10.1007/bf03546509

Sherin, B., and Star, J. R. (2011). "Reflections on the study of teacher noticing," in *Mathematics teacher noticing: seeing throug teachers eyes*. Editors M. Sherin, V. Jacobs, and R. Philipp (New York, NY: Routledge).

Stein, M. K., Engle, R. A., Smith, M. S., and Hughes, E. K. (2008). Orchestrating productive mathematical discussions: five practices for helping teachers move beyond show and tell. *Math. Think. Learn.* 10, 313–340. doi:10.1080/10986060802229675

Steinle, V. (2004). Changes with age in students' misconceptions of decimal numbers. MS dissertation: University of Melbourne. Available at:http://hdl.handle.net/11343/39024.

Stürmer, K., Marczynski, B., Wecker, C., Siebeck, M., and Ufer, S. (in press). "Praxisnahe Lerngelegenheiten in der Lehrerausbildung. Validierung der simulationsbasierten Lernumgebung DiMaL zur Förderung diagnostischer Kompetenz von angehenden Mathematiklehrpersonen." in *Vielfältig herausgefordert. Forschungs- und Entwicklungsfelder der Lehrerbildung auf dem Prüfstand*. Editors N. Beck, T. Bohl, and S. Meissner. (Tübingen: Tübingen University Press).

Südkamp, A., Kaiser, J., and Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: a meta-analysis. *J. Educ. Psychol.* 104 (3), 743–762. doi:10.1037/a0027627

Südkamp, A., Möller, J., and Pohlmann, B. (2008). Der simulierte Klassenraum. Eine experimentelle Untersuchung zur diagnostischen Kompetenz. *Z. für Padagogische Psychol.* 22 (34), 261–276. doi:10.1024/1010-0652.22.34.261

Südkamp, A., and Praetorius, A.-K. (2017). "Eine Einführung in das Thema der diagnostischen Kompetenz von Lehrkräften," in *Diagnostische Kompetenz von Lehrkräften: theoretische und methodische Weiterentwicklungen*. Editors A. Südkamp and A.-K. Praetorius (Münster, Germany: Waxmann), 13–18.

Leuders, T., Leuders, J., and Philipp, K. (Editors) (2018). *Diagnostic Competence of Mathematics Teachers: Unpacking a Complex Construct in Teacher Education and Teacher Practice*. New York: Springer.

Tröbst, S., Kleickmann, T., Depaepe, F., Heinze, A., and Kunter, M. (2019). Transfer from instruction on pedagogical content knowledge about fractions in sixth-grade mathematics to content knowledge and pedagogical knowledge. *Unterrichtswinssenschaft* 47 (1), 79–97. doi:10.1007/s42010-019-00041-y

Tröbst, S., Kleickmann, T., Heinze, A., Bernholt, A., Rink, R., and Kunter, M. (2018). Teacher knowledge experiment: testing mechanisms underlying the formation of preservice elementary school teachers' pedagogical content knowledge concerning fractions and fractional arithmetic. *J. Educ. Psychol.*, 110, 1049. doi:10.1037/edu0000260

Ufer, S., and Neumann, K. (2018). "Measuring competencies," in *International handbook of the learning sciences*. Editors F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, and P. Reimann (New York, NY: Routledge), 433–443.

Van de Pol, J., Volman, M., and Beishuizen, J. (2010). Scaffolding in teacher-student interaction: a decade of research. *Educ. Psychol. Rev.* 22 (3), 271–296. doi:10.1007/s10648-010-9127-6

Van den Kieboom, L. A., Magiera, M. T., and Moyer, J. C. (2013). Exploring the relationship between K-8 prospective teachers' algebraic thinking proficiency and the questions they pose during diagnostic algebraic thinking interviews. *J. Math. Teach. Educ.*, 17, 429. doi:10.1007/s10857-013-9264-1

Wollring, B. (2004). Individualdiagnostik im Mathematikunterricht der Grundschule als Impulsgeber für Fördern, Unterrichten und Ausbildung. Teil 2: Handlungsleitende Diagnostik [Individual diagnosis as catalyst in the primary mathematics classroom. Part 2: Diagnostic for pedagogical decisions] Schulverwaltung. Ausgabe Hessen, Rheinland-Pfalz und Saarland, 8 (11), 297–298.

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

frontiers in Education

CONCEPTUAL ANALYSIS
published: 26 March 2021
doi: 10.3389/feduc.2021.605999

An Argument-Based Framework for Validating Formative Assessment in the Classroom

*Peter Yongqi Gu**

School of Linguistics and Applied Language Studies, Victoria University of Wellington, Wellington, New Zealand

The embedded and contingent nature of classroom-based formative assessment means that validity in the norm-referenced, summative tradition cannot be understood in exactly the same way for formative assessment. In fact, some scholars (e.g., Gipps, Beyond testing: towards a theory of educational assessment, 1994, Falmer Press, London, UK) have even contended for an entirely different paradigm with an independent set of criteria for its evaluation. Many others have conceptualized the validity of formative assessment in different ways (e.g., Nichols et al., 2009, 28 (3), 14–23; Stobart, Validity in formative assessment, 2012, SAGE Publications Ltd, London, UK; Pellegrino et al., Educ. Psychol., 2016, 51 (1), 59–81). This article outlines a framework for evaluating the argument-based validity of CBFA. In particular, I use Kane (J. Educ. Meas., 2013, 50 (1), 1–73) as a starting point to map out the types of inferences made in CBFA (interpretation and use argument) and the structure of arguments for the validity of the inferences (validity argument). It is posited that a coherent and practical framework, together with its suggested list of inferences, warrants and backings, will help researchers evaluate the usefulness of CBFA. Teachers may find the framework useful in validating their own CBFA as well.

Keywords: formative assessment, classroom-based formative assessment, validity, validation of formative assessment, argument-based validation

OPEN ACCESS

Edited by:
Chris Davison,
University of New South Wales, Australia

Reviewed by:
Susan M Brookhart,
Duquesne University, United States
Ricky Lam,
Hong Kong Baptist University, Hong Kong

***Correspondence:**
Peter Yongqi Gu
peter.gu@vuw.ac.nz

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

Received: *14 September 2020*
Accepted: *19 February 2021*
Published: *26 March 2021*

Citation:
Gu PY (2021) An Argument-Based Framework for Validating Formative Assessment in the Classroom. Front. Educ. 6:605999. doi: 10.3389/feduc.2021.605999

INTRODUCTION

Since Black and Wiliam's (1998) review article, formative assessment has gained increasing currency in educational systems as different as Australia (Klenowski, 2011), China (Xu and Harfitt, 2019), New Zealand (Bell and Cowie, 2001), Norway (Hopfenbeck et al., 2015), the United Kingdom (Torrance and Pryor, 1998) and the United States (Ruiz-Primo and Furtak, 2007). Part of the surge of interest comes from its intuitive appeal; part of it comes from claims of its effectiveness in "doubling the speed of student learning" (Wiliam, 2007, 36–37).

Recent years have seen repeated challenges to the effectiveness promise of formative assessment. Dunn and Mulvenon (2009) focused on the lack of consensus on definition. They rightly pointed out that "without a clear understanding of what is being studied, empirical evidence supporting formative evidence will more than likely remain in short supply" (p. 2). Bennet (2011) noted that most of the original claims of effectiveness in Black and Wiliam's (1998) review were exaggerated or misplaced. Kingston and Nash (Kingston and Nash, 2011) did a new meta-analysis of more than 300 studies on the efficacy of formative assessment. They found only 13 studies (42 independent effect sizes) that reported enough information to calculate effect sizes. The average effect size was only 0.20, with formative assessment being more effective in English language arts (effect size = 0.32) than in mathematics (effect size = 0.17) or science (effect size = 0.09). To use Bennet's (2011) words,

the "mischaracterisation" of Black and Wiliam's (1998) conclusions "has essentially become the educational equivalent of urban legend" (p. 12).

It should be noted that none of the challenges denies the potential efficacy of formative assessment. They serve to emphasize a point that formative assessment is not a simplistic issue and that it is not necessarily effective in improving student learning. In addition to different definitions of formative assessment, other factors that influence the effectiveness of formative assessment includes, among others, its domain dependency, teachers' assessment literacy, and support or constraints in the larger educational context.

Most importantly, validity is a necessary but insufficient condition for effectiveness. Even a valid formative assessment task may not lead to intended learning success; invalid formative assessment practices will definitely not be effective. If we follow Kane and Wools (2019) and view validity from both a measurement perspective and a functional perspective, we can reword the previous statement this way: proper assessment procedures and the interpretation and use of assessment results may or may not lead to the functional effect of usefulness. In fact, some forms of formative assessment are more effective than others; and some formative assessment practices may not lead to learning at all. In other words, validating formative assessment is an important step towards ensuring its usefulness.

This article looks at the validity issue of formative assessment, and illustrates how the argument-based framework for test validation (Kane, 2013) can be applied to the validation of formative assessment in the classroom. I will first present an operationalization of classroom-based formative assessment (CBFA), followed by a brief introduction of validity and validation issues in educational measurement in general. Finally, argument-based validation of classroom-based formative assessment will be outlined. I will illustrate how this can be done with a concrete example.

CLASSROOM-BASED FORMATIVE ASSESSMENT

Before we talk about the validity (interpretation and use) of CBFA and its effectiveness, we need to delineate its conceptual boundaries, so that we know exactly what is implemented, summarized as findings, and potentially transferred across contexts (Bennett, 2011). In this section, I will start by operationalizing the construct of formative assessment, and proceed to narrow down the construct into its classroom-based variant. I will also highlight two seminal features as part of this operationalization, i.e., cycle length and a continuum of formality of assessment events, and attempt to locate CBFA as predominantly short-cycle, contingent assessment events that happen in the classroom.

Defining and Operationalizing Formative Assessment

Formative assessment has been understood as instrument, process, and function. The first perspective is in the minority and is represented mostly by test publishers (Pearson Education, 2005). Formative assessment in this sense is reflected in the diagnostic tests they produce. An overwhelming amount of definitions do not view formative assessment as an instrument. Many scholars define formative assessment as a process by which student understanding is elicited and used to adjust teaching and learning (Popham, 2008). Most other definitions see formative assessment as a process aimed at a formative function (Bennett, 2011).

Assessment is formative when evidence of learning is elicited and matched against the learning target to inform the teacher and the learner about the gap between the learner's current state of knowledge or ability and the target. To be helpful at all in closing the gap, a formative assessment event needs to be rounded off with follow-up action (Sadler, 2010). Davison and Leung (2009) outline two basic functions of formative assessment, informing and forming. The former puts emphasis on the necessary but insufficient nature of feedback; while the latter underscores the importance of students' engagement with the feedback they receive in order for learning to take place.

Similarly, Andrade (2010) simply conceptualises formative assessment as "informed action" (p. 345). Expressed in another way, most researchers (Ramaprasad, 1983; Sadler, 1989; Black and Wiliam, 2012) believe that the essence of formative assessment involves establishing 1) where the learners are going; 2) where the learners currently are in their learning; and 3) what needs to be done to get them there.

Formative assessment is hard to operationalize, partly because we normally talk about it being a formative function of assessment rather than a type of assessment with a palpable format. Elsewhere, I have tried to operationalize formative assessment into formative functions and formative practices (Gu, 2020). The former includes a formative purpose before assessment and a formative effect being achieved at the end. The latter includes four crucial consecutive steps: eliciting evidence of learning or understanding, interpreting the evidence, providing feedback, and student/teacher action engaging with the feedback. Each of the four steps is oriented towards achieving a concrete target of learning (**Figure 1**). Ideally, a formative assessment event should include a formative purpose, a formative practice cycle (which I call a formative event), and achieve a formative effect. In most cases, however, we cannot realistically expect to achieve any learning effect with one round of formative practice. Very often we do not have an explicit and conscious formative purpose before we start a round of formative practice inside the classroom. I therefore see one complete round of formative practice involving all four steps moving towards achieving the target of learning as the minimum requirements for the defining features of a formative assessment event. This operationalization allows teachers to catch formative assessment as it appears, as it were, and gives researchers concrete units for analysis (Gu and Yu, 2020).

Classrooms as a major site for learning is a major site for formative assessment as well. However, not all assessment that happens in the classroom is formative. Formative assessment that happens in the classroom can be planned or contingent; and, depending on the task being assessed, classroom-based formative assessment can be completed within short, medium, and long cycles.

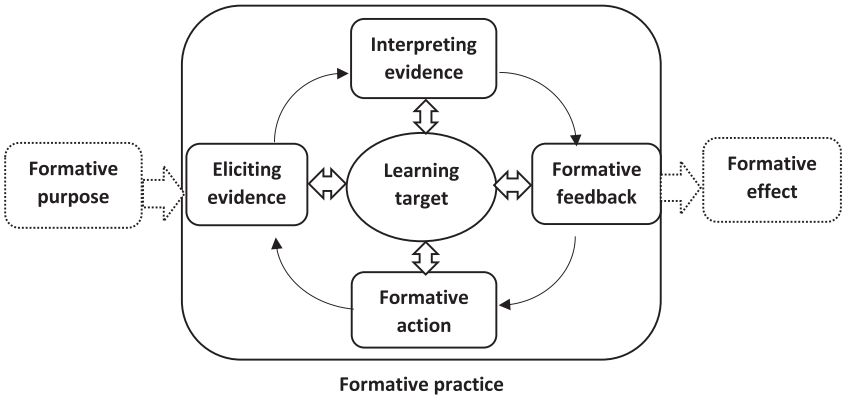


FIGURE 1 | Operationalizing formative assessment.

Delimiting Classroom-Based Formative Assessment

Teachers use a wide range of tools to collect information about student learning in class. Sometimes it can be a formal test; other times it may just be an informal question or an observation of a regular learning task. However, not all classroom tasks are assessment, and classroom-based assessment is not necessarily formative (Black and Wiliam, 2005). Furthermore, formative assessment does not necessarily happen in the classroom. Continuous assessment such as a class quiz, for example, definitely takes the form of an assessment, and it can be done in regular intervals during a period of teaching. Unless the information elicited through the quiz is interpreted and the result relayed back to the students, and unless the students act on the feedback from the quiz result, nothing becomes formative. Likewise, not all alternative forms of assessment such as peer grading can achieve formative functions (Davison and Leung, 2009). Many times, classroom tasks elicit information about student learning that the teacher and the students may not become aware of. Even if this information is noticed by the teacher, and feedback is provided, if the student concerned does not take any action in response to the feedback, the feedback will be wasted.

Classroom-based formative assessment is therefore a teaching/learning event that serves a formative assessment function and which happens within or beyond one class. One complete CBFA event includes 1) elicitation of evidence of students' understanding or learning, 2) interpretation of the elicited information against the learning target or success criteria, 3) feedback based on this interpretation for the student in question, and 4) follow-up action taken by the student and/or teacher to improve learning. All these elements must be present before each CBFA event is complete. And more often than not, learning only takes place after the completion of a series of these cyclical, and spiralling CBFA events.

Cycles of Formative Assessment Events

Classroom assessment practices that involve elicitation of evidence, interpreting the evidence, providing feedback, and student/teacher take-up and action form one complete CBFA event (**Figure 1**). Each event is aimed at a target of learning, teaching, and assessment; and each step or element has the learning target as the reference point. These elements are both sequential and interactive. The completion of one cycle normally will necessitate a readjustment of the target which entails another cycle of assessment practice. The elements, therefore, form spiralling cycles, with each complete cycle moving student understanding or learning closer to the target. This happens continuously until a judgment is made that the target is reached and the success criteria met.

Depending on the scope of the task being assessed, a complete cycle of an assessment event mentioned above can take a few seconds; or it may take a week or much longer to complete. Wiliam (2010) groups the lengths of these cycles into three types: short-, medium-, and long-cycles (**Table 1**).

(Wiliam 2010, 30)

As **Table 1** suggests, CBFA normally belongs to the 'short-cycle' category. This is especially true for those assessments that happen within the classroom. That said, learning usually takes place in timespans longer than a normal class. It is, therefore, often the case that teachers and learners need to check again and again in order to see the effect of learning and see if a course of action works. These actions would take longer than one class and can also be regarded as CBFA. Formative assessment events that go beyond a month or so to complete are normally more formal. For example, information from a formal diagnostic test can be used to guide learning efforts for a whole semester or more. These normally happen well beyond regular classes, and, despite being formative in nature, cannot be counted as CBFA anymore, simply because most of the assessment practices do not happen inside the classroom.

Planned and Contingent Assessment Practices

When formative assessment practices are examined inside the classroom, Cowie and Bell (1999) found largely two types, planned and interactive assessment practices. For planned

TABLE 1 | Short-, medium-, and long-cycle lengths for formative assessment.

Type	Focus	Length
Long-cycle	Across marking periods, quarters, semesters, years	4 weeks to 1 year
Medium-cycle	Within and between instructional units	1–4 weeks
Short-cycle	Within and between lessons	Day by day: 24–48 h minute by minute: 5 s to 2 h

formative assessment, the teacher has a clear but usually general purpose and target before class, s/he deliberately chooses assessment tools to collect information about students' understanding of or performance on the target task, interpret the result on the spot or after class, provides feedback and act on it. A questionnaire before teaching starts would help the teacher gauge the students' current level and expectations, which in turn will help the teacher prepare for more targeted teaching. Likewise, weekly quizzes and many curriculum-embedded tests that are pre-designed for a unit of teaching help the teacher monitor the learning progress of the class and adjust teaching accordingly.

Inside the classroom, many assessment opportunities arise spontaneously without the teacher's preparation. These normally take the form of classroom interactions or the teacher's observations of the students' task performances. Cowie and Bell (1999) labelled these assessment events 'interactive'. Interactive formative assessment events are usually triggered by the teacher noticing an unexpected or erroneous understanding or performance. On the spot interpretation of the deviant understanding would help the teacher recognize the error as a significant point to focus on. The teacher may immediately ask another student the same question and see if the problem is pervasive (both a follow-up action of the previous assessment event and the start of another assessment event), and if the gravity of the problem is deemed serious, the teacher may decide to explain, re-teach, or change a practice activity for the whole class.

The same phenomenon has been observed by Ruiz-Primo and her colleagues who labelled it 'informal formative assessment' (Ruiz-Primo and Furtak, 2006; Ruiz-primo and Furtak, 2007; Ruiz-Primo, 2011). These researchers developed this into an observation framework that included eliciting (E), student response (S), recognizing (R), and using information (U) and called it the 'ESRU cycle'. Interestingly, their studies indicated that informal teacher classroom assessment practices include different configurations in terms of how many elements are practiced. Few complete cycles of informal formative assessment were found. Instead, teachers used ES more often than ESR and ESRU. Those who used more complete ESRU cycles were found to benefit their students better.

Meanwhile, many researchers realize that it is often hard to categorize CBFA events into dichotomies such as planned/unplanned or formal/informal. The dichotomies are in fact two ends of a continuum. Shavelson et al. (2008) outline three anchor points on a continuum: (a) "on-the-fly," (b) planned-for interaction, and (c) formal and embedded in curriculum. Similarly, Bailey and Heritage (2008) also referred to a 'degree of spontaneity' (p. 48) and used 'on the run/in the moment', 'planned for interaction', and 'embedded in curriculum' assessment to describe the continuum. Likewise, Davison (2008) talked about 'a typology of possibilities' which also aligned four types of classroom assessment possibilities along a continuum, ranging from 'in-class contingent formative assessment-while-teaching', 'more planned integrated formative assessment', and 'more formal mock or trial assessments modelled on summative assessments but used for formative purposes', to 'prescribed summative assessments, but results also used formatively to guide future teaching/learning'.

An overwhelming proportion of assessment activities happening in classrooms are contingent, and the cycles are short and often incomplete. The formal, semi-formal, and often curriculum-embedded assessment activities in or out of everyday classes can be used for formative purposes as well.

By nature, formative assessment is meant to support learning. This, however, does not imply that any formative assessment practice will necessarily improve learning. Inside the classroom, many factors influence the validity and the effectiveness of the assessment practice. For example, even if a complete formative assessment event is present, the task being assessed can be irrelevant to the curriculum target being taught and learned. One or even more observations of similar tasks performed by a few students may not be enough to lead to a generalizable conclusion. On the spot interpretations of the evidence of learning may or may not be appropriate. Premature claims can be made about student achievement or ability based on the interpretations. Feedback provided and instructional decisions thereafter can be misguided if the interpretation of learning evidence is inaccurate. In other words, the lack of evidence we discussed previously for the effectiveness of formative assessment can well be due to a lack of validity in the formative assessment that has been studied.

VALIDITY AND VALIDATION

In educational measurement, validity refers to "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests"; while validation is seen "as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use" (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014, 11). In this sense, validity of formative assessment is the plausibility of the interpretations and the appropriateness of the feedback and uses based on the evidence of learning elicited. Validation of formative assessment is the process in which interpretations and uses of formative assessment results are specified, justified and supported.

A number of scholars have tried to examine the validity issue of formative assessment. Gipps (1994) contends that assessment for teaching and learning purposes deserves a completely new paradigm for its evaluation. Instead of terminologies such as validity and reliability that belong to the psychometric tradition, new terms such as Curriculum fidelity, Comparability, Dependability, Public credibility, Context description, and Equity represent a set of criteria better suited to formative assessment. Many other scholars (e.g., Stobart, 2012; Kane and Wools, 2019) seem to have come to the conclusion that a validity framework is appropriate for formative assessment, although the emphases in different facets of this framework and the kinds of interpretations and uses of assessment results are very different from psychometric tests (Pellegrino et al., 2016).

Kane and Wools (2019) distinguished between two perspectives on the validity of assessments: a measurement versus a functional perspective. The former focuses on the accuracy of construct scoring, and the latter focuses on the extent to which the assessment serves its targeted purposes. Kane and Wools (2019) argued that, for classroom assessment, "the functional perspective is of central concern, and the measurement perspective plays a supporting role" (p. 11).

A number of scholars (e.g., Stobart, 2012) take a similar position and have placed their emphasis of validity on the effect or the consequential facet of formative assessment, arguing that a major claim is to lead to the improvement of learning. While I do agree that ideally each formative assessment practice leads to targeted learning results, and that this should be the ultimate criterion to evaluate the validity of formative assessment, I do not see it as practical to expect every formative assessment event to result in desired learning consequences. Very simple and concrete learning tasks such as the correct pronunciation of a word may be achievable at the end of a short cycle of formative assessment practice. Most learning tasks, however, will need a much more complex process of teaching, learning and assessment to be completed.

I contend that the "measurement perspective" is equally important for formative assessment, but the emphasis of formative assessment in such a perspective would be very different from traditional tests. Just like the fundamental importance of the psychometric properties of a test in producing the scores for valid interpretations and uses, the basic properties of a formative assessment event (i.e., eliciting evidence of learning, interpreting the results, providing feedback, and acting on feedback) must be carried out appropriately. I would call this an "assessment perspective", and posit that the accuracy and trustworthiness of the information obtained from formative assessment, the correct interpretations and appropriate uses of assessment results determine to a large extent the usefulness of the formative assessment practice.

Most importantly, accurate interpretations and appropriate uses of assessment results very much depend on the assessor's pedagogical content knowledge (Shulman, 1986) which includes, among other things, the learning and assessment target and the success criteria in reaching the target. This domain-specific understanding of the learning target is a crucial facet of classroom formative assessment that makes or breaks any formative assessment practice (Bennett, 2011). Setting the right assessment goal, choosing appropriate tools to elicit the evidence of learning, interpreting the evidence appropriately, providing the right feedback, and embarking on an informed course of action, every stage of an assessment event can go wrong, if the assessor's understanding of the learning target is inappropriate or faulty. For example, in the formative assessment of language learning in class, the teachers' knowledge of curriculum standards, their beliefs in language competence and language learning, and their understanding of the success criteria in performing the language tasks used to elicit evidence of student learning, are as important as, if not more important than the assessment procedures as such.

VALIDATING CLASSROOM-BASED FORMATIVE ASSESSMENT

The Argument-Based Validation Framework

Over the last 2 decades or so, a validation framework that allows all evidences to be presented as a coherent whole (as opposed to a list of fragmented evidences) is getting increasingly accepted by the educational assessment community. The framework is called "argument-based validity". The idea is: in claiming that our assessment is good for its purposes, we are making an argument. Validation is therefore a matter of making this argument convincing enough for people who care about our assessment.

As early as the 1980s, Cronbach (1988) began to see test validation as gathering evidence to support an argument for our design, interpretation, and use of a test. Over the years, Kane (1992), Kane (2001), Kane (2006) and Mislevy et al. (2003) have developed the argument-based approach to test validation into a coherent and practical framework. In language assessment, Bachman (2005) and Bachman and Palmer (2010) have taken up the approach; and one of the major English language tests, TOEFL, has been validated using the argument-based approach (Chapelle et al., 2008). The latest addition is Chapelle's (2020) book-length volume on argument-based validation of language tests.

In an argument-based framework, validation is done in two steps, or to put it another way, we need two sequential arguments to validate an assessment: an interpretation and use argument (IUA) and a validity argument (Kane, 2013). In step 1, we articulate an IUA through a logical analysis of the chain of inferences linking test performance to a judgement or decision, and the assumptions on which they rest. In other words, we outline explicitly the major inferences and claims we are making based on assessment outcomes. In step 2 (validity argument), we provide an overall evaluation of the inferences in the IUA and systematically argue that each claim or inference is true unless proven otherwise. The validity argument uses Toulmin's (2003) argument structure. **Figure 2** shows a simple claim using the Toulmin structure. Since the rebuttal does not overturn the conclusion, the claim stands.

Hopster-den Otter et al. (2019) proposed an argument-based framework to validate formative assessment. They conceptualised

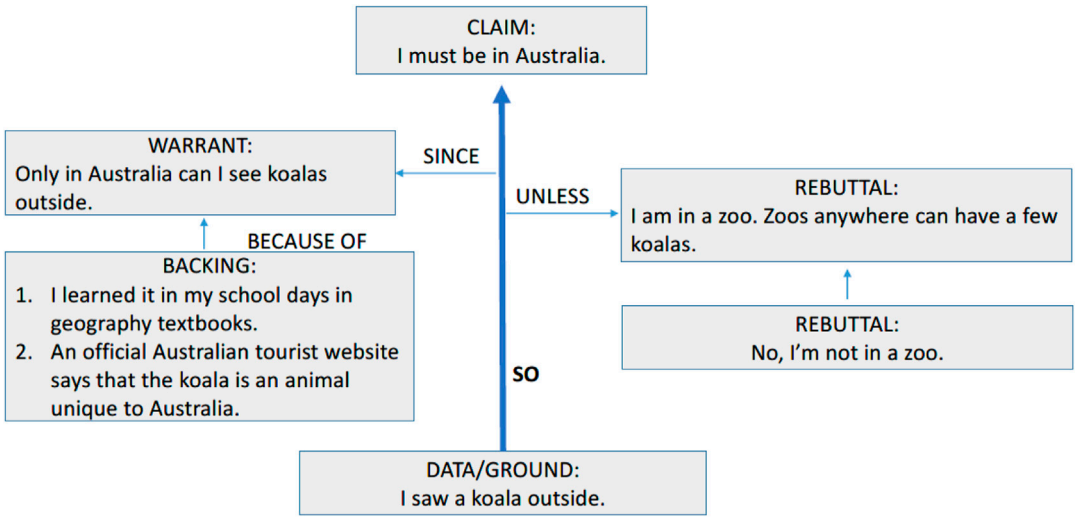


FIGURE 2 | Argument structure for a simple claim/conclusion.

formative assessment as "both an instrument and a process, whereby evidence is purposefully gathered, judged, and used by teachers, students, or their peers for decisions about actions to support student learning" (p. 3). This conceptualisation was confined to the curriculum-embedded, pre-defined types of formal assessment tasks (instruments) that resembled summative tests in format, and excluded the majority of classroom-based formative assessments which occur contingently and unplanned. This explains why their "interpretation inferences" in the IUA being identical to those in tests, which is in line with their previous thinking on "formative use of test results" (Hopster-den Otter et al., 2017).

A major contribution of Hopster-den Otter et al. (2019) lies in their conceptualisation of the Use component of the IUA, focusing on the utilisation of test results for instructional purposes. They parsed the use component of IUA into four inferences: Decision, Judgment, Action, and Consequence. These inferences at the end of a diagnostic test make the use of the instrument formative. In their illustrative example, Hopster-den Otter et al. (2019) referred to the validation of an online test of arithmetic which provided subsequent feedback for primary school teachers and learners.

Seeing formative assessment as formative use of tests necessitates the judgment and use of assessment information after a test. However, conceptualising CBFA as both a process and a function but not an instrument (**Figure 1** above) means that most of the judgment, interpretation, and action after feedback are done during the classroom assessment process. As a result, the validation process in CBFA does not have to start after assessment is done; and the Use component of IUA does not need to be parsed the way Hopster-den Otter et al. (2019) did. In other words, the framework presented next is an alternative to Hopster-den Otter et al. (2019) that complements their framework. While the Hopster-den Otter et al. framework is more appropriate for formative use of tests, the framework in this article is more appropriate for CBFA.

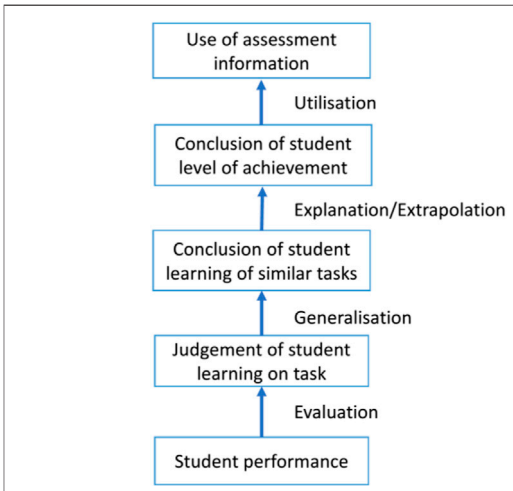


FIGURE 3 | Chain of inferences in CBFA.

Argument-Based Validation of CBFA

Step 1: Interpretation and use argument

The following figure (**Figure 3**) outlines the chain of inferences in CBFA. When we make a judgment of a student's ability in performing a task in class, we are making an evaluation inference. When we conclude that the student is able to do similar tasks across similar situations, we are making a generalization inference. After a number of observations of successful performance on similar tasks, we say that the student has achieved a curriculum criterion (extrapolation), or the student is able to do certain things with language represented by his ability to complete future tasks of a similar nature (explanation). Here we are making two types of the extrapolation inference (extrapolation and explanation). When we use this information to make decisions about this student (e.g.,

TABLE 2 | Claims and inferences in CBFA.

CBFA Claims	Inference links
Claim 1: CBFA judgment is carried out appropriately	Evaluation: linking performance to judgment
Claim 2: CBFA judgment about student achievement is trustworthy	Generalisation: linking individual observation to generalised judgement over all possible observations
Claim 3: CBFA reflects students' expected language achievement	Explanation: linking judgment to interpretation against theoretical construct Extrapolation: linking judgment to interpretation against curriculum targets and teaching
Claim 4: CBFA is used to improve learning outcomes	Utilisation: linking interpretation to use

TABLE 3 | Warrants and their backing in CBFA.

Inference	Assumptions (warrants)	Evidence (backing)
Evaluation	Assessment targets and success criteria are clear; Elicitation tools appropriately chosen and used; and key procedures (elicitation, interpretation, feedback, action) of CBFA have been followed	Interviews of teacher and students to see their understanding of assessment targets and success criteria; Classroom discourse analysis to see assessment types and how they are carried out; and content analysis of classroom recordings to see how elicitation and interpretation are done, what feedback is provided, and what action is taken after feedback.
Generalisation	Classroom performance on language tasks is consistent across similar tasks, assessors, assessment forms and occasions	Multiple sources of evidence; Multiple observations; Sample observation tasks are representative of content domain tasks; and sample observation conditions are representative of content domain conditions
Explanation	Classroom assessment tasks engage the same abilities and processes as those in the theoretical construct of language competence appropriate for the context of teaching	Checking construct relevance and construct representativeness Interviews; Observation of assessment processes; Discourse/conversation analysis; and logical analysis of assessment tasks
Extrapolation	Assessment tasks and materials are representative of the knowledge, skills, and abilities targeted by the curriculum at the relevant level (content domain)	Judgmental evidence that assessment tasks are representative samples of the content domain; and logical analysis of assessment task content
Utilisation	Information provided to users are useful and sufficient (informing); and assessment information is used to adjust learning and teaching (forming)	Analysis of feedback (type, informativeness); Analysis of adjustment to learning and teaching; Analysis of adjustment to learning and teaching; Improved score in exams

he can go to the next level; or he needs more efforts to improve on this standard), we are making a utilization inference.

Since most assessment tasks in CBFA are contingent classroom activities, the assessor (mostly the teacher) makes judgements and decisions on the spot and does not wait till the end of the activity to interpret evidences of student learning. These explanation and extrapolation inferences and the judgements and feedback are much more closely bundled together than those a teacher makes at the end of a test. In addition, since the conceptualisation of CBFA in this framework does not assume formative effects being achieved, for the sake of parsimony, the Utilisation inference in the proposed IUA chain is not further parsed into sub-inferences.

Table 2 elaborates on the four major claims of classroom-based formative assessment. These four claims and their associated inferences make up the interpretation and use argument (IUA).

Step 2: Validity argument

After the articulation of the IUA, the next step is to argue with supporting reasons or warrants that all claims and inferences are plausible. In many cases, we also need to prove that alternative reasoning (rebuttal) is not supported by evidence; otherwise our claims will not stand if evidences are found to back up the rebuttals. **Table 3** lists the warrants and their potential backings for the validity argument of CBFA.

Argument-Based Validation of CBFA: An Example

Let's now look at a CBFA event, and see how it can be evaluated using the argument-based approach. Due to a lack of space, I will be deliberately short, and will not be illustrating all the details in the two-step validation process.

The following classroom assessment event forms a complete assessment cycle and should be counted as CBFA. Is this CBFA good enough for its intended purpose?

For the interpretation and use argument, I have largely indicated the list of inferences and claims for this CBFA event, although the wording is not in the format of a claim or inference. The IUA is illustrated in **Figure 4**.

Validity argument should next be provided for each of the above claims. I will take the explanation claim and show that it is not true (**Figure 5**). In other words, the teacher's interpretation of the assessment outcome is wrong. In these cases, no matter how useful the follow-up actions are, they will not help solve the targeted learning problem, thus not achieving the effect of CBFA.

- We had an in-class shared reading task today. I went around class and observed the students. My observation focused on three groups and I found a number of problems in understanding (evaluation).
- I realized that many students couldn't understand this type of reading (generalization).
- The students' lack of vocabulary is a concern (explanation).
- I told them they needed a larger vocabulary to become better readers; and assigned them a task to memorize 50 words a week from now on (utilization).

After generalising classroom observations of students' reading problems, the teacher could have arrived at the conclusion that the students/class were not achieving a particular curriculum target of reading, or that they would have problems reading similar texts in real world tasks (Extrapolation). She could have also inferred that this evidence in class revealed the students' deficiency or imperfect learning in certain areas of reading competence (Explanation). The teacher opted for the latter but identified a wrong component (vocabulary size) of the construct of reading as the cause of the problem in the Interpretation phase of this CBFA. While the transient nature of many CBFA events would make it unavoidable for some wrong interpretations of assessment data, this example illustrates the importance of teacher pedagogical content knowledge, a crucial aspect of assessment literacy that makes or breaks a formative assessment decision.

The CBFA cycle in this example may take slightly longer than normal to complete, because the action component comes after class. While the consequential aspect of the formative assessment cycle can only become possible after a full round, validity argument for each inference can be done any time during the whole spiralling process. This validity argument during the process as soon as an inference is made explicit in an IUA is a key part of the formative mechanism that makes flexible adjustment of teaching and learning possible. In the example, exercises in explicating the IUA inferences (**Figure 4**) make teachers more aware of their own decision-making processes in making use of assessment during instruction. Likewise, a validity argument (**Figure 5**) for each inference will help teachers decide whether and what changes are needed to achieve the formative effect. Without the validity argument, for example, the students may go on following the teacher's advice to remember more vocabulary items, and the real problem of reading identified at the elicitation stage may never been dealt with.

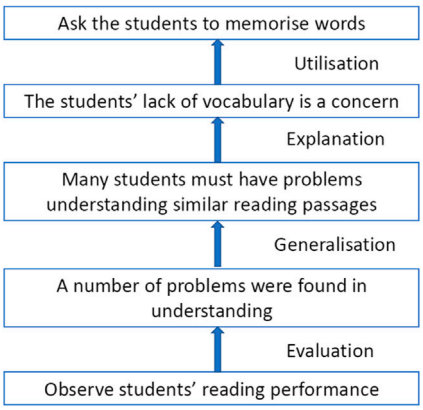


FIGURE 4 | Interpretation and use argument.

Who does CBFA validation, when, how?

Ideally, teachers themselves should validate their own CBFA as and when it happens in class. Teachers should also form communities of assessment practice in and beyond their own schools, so that peer teachers can help each other validate their CBFA. In addition, university researchers should join these communities of assessment practice every now and then to bring further theoretical and empirical expertise and to oversee that CBFA is done appropriately.

Both planned and contingent CBFA should be validated as often as needed, in any case, regularly. After all, as we have seen, despite its powerful potential, CBFA is only as good as the way it is used in class. Informal validation of CBFA should happen as and when it occurs in class. Formal validation can take the form of peer moderations and class observations. Teachers can also video-record their own classes for formal analysis at a later time. In the example above, the wrong interpretation of CBFA evidence could have been caught if the teacher or a peer validated her CBFA practices by going through her own video data of the lesson. She could then reinterpret the evidence available, and provide other alternatives of potential action in future classes. In addition, lesson plans can also be analysed for planned assessment practices and potential contingent CBFA.

CONCLUSION

In this article, I have offered an operational definition of formative assessment and classroom-based formative assessment. I argued that a clear operationalisation is the starting point for researchers and teachers alike to examine the validity and effectiveness of the formative assessment construct. Next, I contended that formative assessment is not necessarily useful in bringing about the desired formative effect, and that validation is needed for even informal and contingent classroom-based assessment events.

The argument-based approach to validation was next introduced. This includes two steps, an explication of the inferences we make from the assessment results followed by an argument for or against each inference using the Toulmin structure of argumentation. In other words, assessment validation is seen as systematically arguing that the interpretations and uses of assessment results are backed up by evidence and theory.

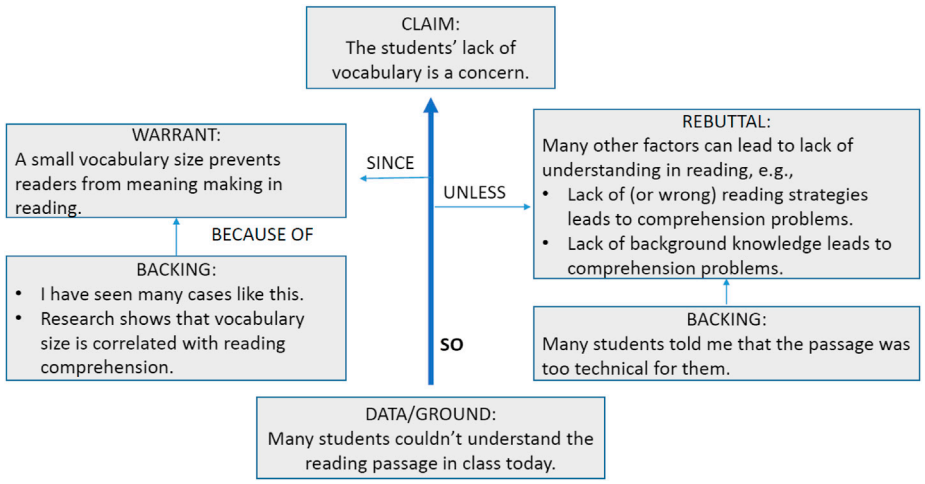


FIGURE 5 | The claim does not stand. Note that this is almost a "slow motion" replay of the argument-based validation process. In real life situations, CBFA validation can be done very quickly and informally without listing all the inferences on paper or drawing up the argument tree of warrants and backings.

Finally, I used an example from an English as a foreign language teacher's CBFA practice to illustrate how validation of CBFA can take place and how overturning one claim can invalidate the overall CBFA inference chain. The article finished by calling for more validations of CBFA not just for research purposes but also for teaching and teacher professional development purposes as well.

A clear operational definition will help teachers implement formative assessment inside their classrooms. A coherent and workable validation framework can assist teachers monitor and evaluate the interpretations and uses of their CBFA practices. This article points to a direction in which CBFA can be validated so that it achieves the formative effect of improved learning.

In using the proposed validation framework, we need to remind ourselves that validation is an ongoing process and that validity is not an either/or concept. Different CBFA events will show different degrees of validity when we go through a validation process. The more confident we are about our assessment outcomes and their interpretations and uses, the more likely we will achieve our intended formative effects.

The validation framework can also be seen as a useful tool for teacher learning. When teachers perform the acts of validation, they will immediately realise that the IUAs are mini-theories in their minds. These mini-theories include the set of criteria teachers make use of on the spot: explicit, latent, and meta-criteria (Sadler, 1985; Wyatt-Smith and Klenowski, 2013) about the nature of the knowledge or competence being assessed and about the criteria for success; they also include the teacher's understanding of how the knowledge is best learned or taught. These mini-theories guide the teacher's interpretation and use of the evaluative task. The more teachers perform validation of their own CBFA practices, the more they become aware of the adequacy of their pedagogical content knowledge behind their assessment. In this sense, validation practices as outlined in this article can also serve as a tool for teacher professional development.

AUTHOR CONTRIBUTIONS

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

REFERENCES

American Psychological Association, and National Council on Measurement in Education (2014). in *Standards for educational and psychological testing* (Washington, D.C.: American Educational Research Association).

Andrade, H. L. (2010). "Summing up and moving forward: key challenges and future directions for research and development in formative assessment," in *Handbook of formative assessment*. Editors H. L. Andrade and G. J. Cizek (New York, NY: Routledge), p. 344–351.

Bachman, L. F., and Palmer, A. (2010). *Language assessment in practice: developing Language Assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Lang. Assess. Q.* 2 (1), 1–34. doi:10.1207/s15434311laq0201_1

Bailey, A. L., and Heritage, M. (2008). *Formative assessment for literacy, grades K-6: building reading and academic language skills across the curriculum.* Thousand Oaks, CA: Corwin.

Bell, B., and Cowie, B. (2001). *Formative assessment and science education*. New York, NY: Springer.

Bennett, R. E. (2011). Formative assessment: a critical review. *Assess. Educ. Principles Pol. Pract.* 18 (1), 5–25. doi:10.1080/0969594x.2010.513678

Black, P., and Wiliam, D. (1998). Assessment and classroom learning. *Assess. Educ. Principles Pol. Pract.* 5 (1), 7–74. doi:10.1080/0969595980050102

Black, P., and Wiliam, D. (2005). Classroom Assessment is not (necessarily) formative assessment (and vice-versa). *Yearbook Natl. Soc. Study Educ.* 103 (2), 183–188. doi:10.1080/0969595980050102

Black, P., and Wiliam, D. (2012). "Developing a theory of formative assessment," in *Assessment and learning*. Editor J. Gardner. 2nd ed. (London, UK: SAGE Publications Ltd), p. 206–230.

Chapelle, C. A., Enright, M., and Joan, J. (2008). *Building a validity argument for the test of English as a Foreign Language*. New York, NY: Routledge.

Chapelle, C. A. (2020). *Argument-based validation in testing and assessment*. Thousand Oaks, CA: SAGE Publications, Inc.

Cowie, B., and Bell, B. (1999). A model of formative assessment in science education. *Assess. Educ. Principles Pol. Pract.* 6 (1), 101–116. doi:10.1080/09695949993026

Cronbach, L. J. (1988). "Five perspectives on validity argument," in *Test Validity Howard wainer and Henry I. Braun, 3–17* (Hillsdale, NJ: Lawrence Erlbaum Associates).

Davison, C., and Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Q.* 43 (3), 393–415. doi:10.1002/j.1545-7249.2009.tb00242.x

Davison, C. (2008). *Assessment for learning: building inquiry-oriented assessment communities*. New York, N.Y.: Routledge.

Hopster den Otter, D. H., Wools, S., Eggen, H. M. J. T., and Veldkamp, B. P. (2019). A general framework for the validation of embedded formative assessment. *J. Educ. Meas.* 56 (4), 715–732. doi:10.1111/jedm.12234

Dunn, K. E., and Mulvenon, S. W. (2009). A critical review of research on formative assessments: the limited scientific evidence of the impact of formative assessments in education. *Pract. Assess. Res. Eval.* 14 (7), 241. doi:10.4324/9780203462041_chapter_1

Gipps, C. V. (1994). *Beyond testing: towards a theory of educational assessment*. London, UK: Falmer Press.

Gu, P. Y., and Yu, G. (2020). Researching classroom-based assessment for formative purposes. *Chin. J. Appl. Linguist.* 43 (2), 150–168. doi:10.1515/cjal-2020-0010

Gu, P. Y. (2020). *Classroom-based formative assessment*. Beijing, BJ: Foreign Language Teaching and Research Press.

Hopfenbeck, T. N., Flórez Petour, M. T., and Tolo, A. (2015). Balancing tensions in educational policy reforms: large-scale implementation of assessment for learning in Norway. *Assess. Educ. Principles Pol. Pract.* 22 (1), 44–60. doi:10.1080/0969594x.2014.996524

Hopster-den Otter, D., Wools, S., Eggen, H. M. J. T., and Veldkamp, B. P. (2017). Formative use of test results: a user's perspective. *Stud. Educ. Eval.*, 52, 12–23. doi:10.1016/j.stueduc.2016.11.002

Kane, M. T., and Wools., S. (2019). "Perspectives on the validity of classroom assessments," in *Classroom Assessment and educational measurement*. Editors S. M. Brookhart and J. H. McMillan (New York, NY: Routledge), p. 11–26.

Kane, M. T. (1992). An argument-based approach to validity. *Psychol. Bull.* 112 (3), 527–535. doi:10.1037/0033-2909.112.3.527

Kane, M. T. (2001). Current concerns in validity theory. *J. Educ. Meas.* 38 (4), 319–342. doi:10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2006). "Validation," in Educational measurement. Editors R. L. Brennan 4th ed. (Westport, CT: American Council on Education/Praeger), p. 17–64.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *J. Educ. Meas.* 50 (1), 1–73. doi:10.1111/jedm.12000

Kingston, N., and Nash, B. (2011). Formative assessment: a meta-analysis and a call for research. *Educ. Meas. Issues Pract.* 30 (4), 28–37. doi:10.1111/j.1745-3992.2011.00220.x

Klenowski, V. (2011). Assessment for learning in the accountability era: queensland, Australia. *Stud. Educ. Eval.* 37 (1), 78–83. doi:10.1016/j.stueduc.2011.03.003

Mislevy, R. J., Steinberg, L. S., and Almond, R. G. (2003). Focus article: on the structure of educational assessments. *Meas. Interdiscip. Res. Perspec.* 1 (1), 3–62. doi:10.1207/s15366359mea0101_02

Nichols, P. D., Meyers, J. L., and Burling, K. S. (2009). A framework for evaluating and planning assessments intended to improve student achievement. *Educ. Meas. Issues Pract.* 28 (3), 14–23. doi:10.1111/j.1745-3992.2009.00150.x

Pearson Education. (2005). Achieving student progress with scientifically based formative assessment White paper, Pearson Education Ltd. Available at: http://www.pearsoned.com/wp-content/themes/pearsoned.com_legacy/pdf/RESRPTS_FOR_POSTING/PASeries_RESEARCH/PA1.%20Scientific_Basis_PASeries%206.05.pdf (Accessed September 6, 2014).

Pellegrino, J. W., DiBello, L. V., and Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educ. Psychol.* 51 (1), 59–81. doi:10.1080/00461520.2016.1145550

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

Ramaprasad, A. (1983). On the definition of feedback. *Syst. Res.* 28 (1), 4–13. doi:10.1002/bs.3830280103

Ruiz-primo, M. A., and Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *J. Res. Sci. Teach.* 44 (1), 57–84. doi:10.1002/tea.20163

Ruiz-Primo, A. M., and Furtak, E.M. (2006). 'Informal formative assessment and scientific inquiry: exploring teachers' practices and student learning'. *Educ. Assess.* 11 (3), 205–235.

Ruiz-Primo, M. A. (2011). Informal formative assessment: the role of instructional dialogues in assessing students' learning. *Stud. Educ. Eval.* 37 (1), 15–24. doi:10.1016/j.stueduc.2011.04.003

Sadler, D. R. (1985). The origins and functions of evaluative criteria. *Educ. Theor.* 35 (3), 285–297. doi:10.1111/j.1741-5446.1985.00285.x

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instr. Sci.* 18 (2), 119–144. doi:10.1007/bf00117714

Sadler, D. R. (2010). Beyond feedback: developing student capability in complex appraisal. *Assess. Eval. Higher Educ.* 35 (5), 535–550. doi:10.1080/02602930903541015

Shavelson, R. J, Donald, B. Y., Carlos, C. A., Paul, R. B., Erin, M. F., Maria, A. R. P., et al. (2008). On the impact of curriculum-embedded formative assessment on learning: a collaboration between curriculum and assessment developers. *Appl. Meas. Educ.* 21, 295–314.

Shulman, L. S. (1986). Those who understand: knowledge growth in teaching. *Educ. Res.* 15 (2), 4–14. doi:10.3102/0013189x015002004

Stobart, G. (2012). "Validity in formative assessment," in *Assessment and learning*. Editor J. Gardner. (London, UK: SAGE Publications Ltd), p. 133–146.

Torrance, H., and Pryor, J. (1998). *Investigating formative assessment: teaching, learning and assessment in the classroom*. Maidenhead, UK: Open University Press.

Toulmin, S. E. (2003). *The uses of argument*. 2nd edn. New York, NY: Cambridge University Press.

Wiliam, D. (2007). Changing classroom practice. *Educ. Leadersh.* 65 (4), 36–42.

Wiliam, D. (2010). "An integrative summary of the research literature and implications for a new theory of formative assessment," *in Handbook of formative assessment. H. L. Andrade and G. J. Cizek (New York, NY: Routledge),* p. 18–40.

Wyatt-Smith, C., and Klenowski, V. (2013). Explicit, latent and meta-criteria: types of criteria at play in professional judgement practice. *Assess. Educ. Principles Pol. Pract.* 20 (1), 35–52. doi:10.1080/0969594x.2012.725030

Xu, Y., and Harfitt, G. (2019). Is assessment for learning feasible in large classes? Challenges and coping strategies from three case studies. *Asia-Pac. J. Teach. Educ.* 47 (5), 472–486. doi:10.1080/1359866x.2018.1555790

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.

frontiers in Education

ORIGINAL RESEARCH
published: 01 April 2021
doi: 10.3389/feduc.2021.628100

Pre-service Teachers' Decision-Making and Classroom Assessment Practices

Cherry Zin Oo[1], Dennis Alonzo[2] and Chris Davison[2]*

[1] Department of Educational Psychology, Yangon University of Education (YUOE), Yangon, Myanmar, [2] School of Education, University of New South Wales, Kensington, NSW, Australia

Classroom assessment practices play a pivotal role in ensuring effective learning and teaching. One of the most desired attributes of teachers is the ability to gather and analyze assessment data to make trustworthy decisions leading to supporting student learning. However, this ability is often underdeveloped for a variety of reasons, including reports that teachers are overwhelmed by the complex process of data analysis and decision-making and that often there is insufficient attention to authentic assessment practices which focus on assessment *for* learning (A*f*L) in initial teacher education (ITE), so teachers are uncertain how to integrate assessment into teaching and make trustworthy assessment decisions to develop student learning. This paper reports on the results of a study of the process of pre-service teachers' (PSTs) decision-making in assessment practices in Myanmar with real students and in real classroom conditions through the lens of teacher agency. Using a design-based research methodology, a needs-based professional development program for PSTs' assessment literacy was developed and delivered in one university. Following the program, thirty PSTs in the intervention group were encouraged to implement selected assessment strategies during their practicum. Semi-structured individual interviews were undertaken with the intervention group before and after their practicum in schools. This data was analyzed together with data collected during their practicum, including lesson plans, observation checklists and audiotapes of lessons. The analysis showed that PSTs' decision-making in the classroom was largely influenced by their beliefs of and values in using assessment strategies but, importantly, constrained by their supervising teachers. The PSTs who understood the principles of A*f*L and wanted to implement on-going assessment experienced tension with supervising teachers who wanted to retain high control of the practicum. As a result, most PSTs could not use assessment strategies effectively to inform their decisions about learning and teaching activities. Those PSTs who were allowed greater autonomy during their practicum and understood A*f*L assessment strategies had greater freedom to experiment, which allowed them multiple opportunities to apply the result of any assessment activity to improve both their own teaching and students' learning. The paper concludes with a discussion of the kind of support PSTs need to develop their assessment decision-making knowledge and skills during their practicum.

Keywords: teacher decision-making, assessment practices, assessment *for* learning, pre-service teacher, teacher agency, initial teacher education, practicum

OPEN ACCESS

Edited by:
Susan M. Brookhart, Duquesne University, United States

Reviewed by:
Peter Ralph Grainger, University of the Sunshine Coast, Australia
Eric C. K. Cheng, The Education University of Hong Kong, Hong Kong

****Correspondence:***
Cherry Zin Oo
cherryzinoo@yuoe.edu.mm

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

Received: *11 November 2020*
Accepted: *15 March 2021*
Published: *01 April 2021*

Citation:
Oo CZ, Alonzo D and Davison C (2021) Pre-service Teachers' Decision-Making and Classroom Assessment Practices. Front. Educ. 6:628100. doi: 10.3389/feduc.2021.628100

INTRODUCTION

Teacher decision-making is essential for effective learning and teaching. A range of research studies highlight the impact of teacher decision-making process on improving student learning (McMillan, 2003; Mccall, 2018; van Phung, 2018). Teachers' analysis of student data helps to reveal students' learning needs, which can then be addressed by implementing appropriate learning interventions, highlighting the importance of evidence-informed teacher decision-making skills (McMillan, 2003). To translate these skills into actual student learning gains, there is a need to ensure that teachers are confident and well-equipped to gather and analyze assessment data to make trustworthy decisions leading to supporting student learning.

However, previous research has highlighted that teachers often struggle to justify their use of assessment approaches (Brookhart, 1991; McMillan, 2001; van Phung, 2018). Many report feeling overwhelmed by the complexity of data analysis and decision-making (McMillan, 2003). As teachers' decision-making is intrinsically a social and cultural experience (Klenowski, 2013), it can be studied through the lens of teacher agency, that is, analyzing how teachers respond to emerging situations in their environment (Priestley et al., 2013, 2015). In teacher decision-making, teacher agency is influenced by the interaction of the context, factors within the school, and the individual teachers' beliefs and values (Priestley et al., 2015).

In the area of assessment decision-making, most published research concerns the nature of teacher decision-making in marking, grading, and high-stakes testing (McMillan and Nash, 2000; Bowers, 2009; Cheng and Sun, 2015; Kippers et al., 2018). However, as the focus of assessment policy has shifted from summative assessment (assessment *of* learning) to formative assessment (assessment *for* learning) (Assessment Reform Group, 2002), more research into teacher decision-making in formative assessment situations is needed. Mccall (2018) suggests that further studies need to be carried out to explore teacher assessment decision-making process, especially in relation to assessment *for* learning (A*f*L) and formative assessment practices. However, such teacher decision-making requires far more than a knowledge and understanding of measurement concepts (McMillan, 2003); it requires new forms of teacher assessment literacy (Alonzo, 2016; Davison, 2019). This study uses Alonzo's (2016) concept of teacher A*f*L literacy anchored to the principles of A*f*L, that is "the knowledge and skills to make highly contextualized, fair, consistent and trustworthy assessment decisions to inform learning and teaching to effectively support both students and teachers' professional learning (p. 58)."

Teachers need to be skilled and knowledgeable in A*f*L practices before they enter their profession, so that they can decide which assessment strategies are best used to improve student learning. The problem is that much research has shown that pre-service teachers (PSTs) are not always well-prepared in initial teacher education (ITE) to use appropriate assessment strategies to support student learning (Volante and Fazio, 2007; Siegel and Wissehr, 2011; Vogt and Tsagari, 2014; BOSTES, 2016). A theoretical introduction to the basic concepts of assessment in a course is inadequate support to be literate in assessment (Popham, 2011; Greenberg and Walsh, 2012). As a result, PSTs do not have enough confidence in applying assessment knowledge and building their skills (Ogan-Bekiroglu and Suzuk, 2014). Therefore, PSTs need to be given the opportunity to apply understandings in classroom practices, including building effective assessment practices (Grainger and Adie, 2014; McGee and Colby, 2014; DeLuca and Volante, 2016).

This paper reports on a study which investigated the ways in which PSTs made classroom assessment decisions with real students and in real classroom conditions whilst undertaking their final practicum. The study addressed the following research questions:

(1) What factors influence PSTs' assessment decision-making processes?
(2) How do these factors facilitate or constrain PSTs' assessment decision-making?

THEORETICAL FRAMEWORK

Teacher agency (Priestley et al., 2013, 2015) was chosen as a framework for this study. This perspective on agency is grounded in the sociology and philosophy of action. Teacher agency determines how teachers respond to emerging situations in their environment, resulting from "the interplay of individual efforts, available resources and contextual and structural factors as they come together in particular and, in a sense, always unique situations" (Biesta and Tedder, 2007, p. 137). Teacher agency is the outcome of the interplay of three dimensions: iterational (teachers' past habitual personal and professional experience); projective (orientation to the future); and practical-evaluative (engagement with cultural, structural, and material context). Teacher agency was used in researching one A*f*L strategy, rubrics, by Heck (2020) who highlighted the role of agency in improving academics' assessment literacy and practice. This study uses teacher agency to help explain how PSTs develop their decision-making skills in terms of using assessment strategies to support student learning.

Teacher agency can be achieved by engaging with the available resources, and contextual elements in school (Stritikus, 2003), enabling PSTs to make decisions about what assessment strategies to use by drawing on from the results of interactions of these three dimensions. van der Nest et al. (2018), who studied the impact of formative assessment activities on the development of teacher agency, argue that agency is the outcome of teachers' engagement with their environment, influenced by their past experience and guided by their future orientation. Individual agency depends on the extent of engagement in the process of learning (Billett, 2004), however, teacher agency is reliant on negotiated assessment procedures (Verberg et al., 2016).

PRE-SERVICE TEACHERS' DECISION-MAKING PROCESS AND THEIR ASSESSMENT PRACTICES

Building PST capacity for assessment decision-making before entering the profession is crucial in ITE. Piro et al. (2014) argues that the curricula of teacher education programs should support PSTs to build their decision-making based on student assessment data. They describe the effective use of an intervention that teaches PSTs how to work with assessment data in ITE. However, Piro's study focused only on using summative assessment data such as standardized testing and end-of-course assessment data for accountability purposes. Similarly, Cramer et al. (2014) looked at PSTs' decision-making based on the use of summative assessment data rather than on data to be used for formative assessment purposes. Therefore, preparing PSTs for effective decision-making should move beyond summative assessment to engage with formative assessment purposes.

A closer look at assessment data intervention studies in ITE shows the need for authentic classroom practices to improve assessment decision-making of PSTs. For example, Reeves and Chiang (2018) explored the effectiveness of data literacy intervention for both in-service and PSTs. Although assessment data practices are embedded in in-service teachers' intervention, assessment practices for PSTs are still limited. Piro and Hutchinson (2014) and Reeves and Honig (2015) included student assessment data that PST could work with, however, A*f*L is an ongoing activity where teachers need to draw on a range of different resources in their decision-making about assessment, including interaction with their students.

The work of Black and Wiliam (1998b) and Hattie (2008) highlight that preparing teachers to be literate in assessment, particularly the use of A*f*L has the highest potential to increase students outcomes. Assessment courses provided in ITE can be classified into three different types: stand-alone assessment courses that are heavily weighted toward theoretical assessment principles, assessment courses including assessment tasks using real students' work, and assessment courses including real assessment practices. To prepare classroom-ready teachers effectively in assessment, they need this last kind of course, with practical opportunities to improve their learning by reflecting on how to apply key assessment principles to help students (Hill et al., 2013; BOSTES, 2016) in order to make trustworthy assessment decisions that help students improve.

ITE programs need to ensure that PSTs have adequate A*f*L literacy and have provided student teachers with the opportunity to critique existing assessment knowledge and skills. Also, student teachers need to be provided with a range of opportunities to apply this assessment knowledge to actual classroom settings to see the link between theory and practice (Willis, 2007) and make sense of how assessment literacy influences practice. Without practice in real classrooms with real students, PSTs are likely to "replicate more traditional, unexamined assessment practices" (Graham, 2005, p. 619). Therefore, rather than simply teaching them how to collect assessment information, PSTs need to have a chance to work with real students (Davison, 2015)

Practicum experiences have been found to have a positive effect on PST practices and help to identify professional development needs (Heck et al., 2020), although only a handful of studies have investigated the assessment practices of PSTs in their practicum. For example, Xu and Brown (2016) highlight that PSTs need to have enough practice to be able to apply and evaluate their conceptions of assessment, but in their review of studies on teacher assessment literacy from 1985 to 2015, found less than 20 studies addressing the understanding and development of teacher assessment literacy in practice (see also Campbell, 2013; Hill and Eyers, 2016).

In Myanmar, the Basic Education Curriculum framework is an on-going reform introduced in 2015. However, the types of assessments in this framework are still heavily weighted toward examinations such as end of term, end of year exams, and national level assessment (examinations). Classroom-level assessment/school-based assessment grounded in A*f*L is included as a small portion of the whole academic year. As a result, students focus on rote learning to get high marks in their exams (Tin, 2000; Aung et al., 2013; Metro, 2015; Maber et al., 2018), and teachers use tests as practice for the final examination. This reliance on mock tests or old questions from national exams shows how the exam-dominated system encourages students to memorize and recite facts. Due to the pressure this puts on the students to have higher outcomes, after-school classes (private tuition) are proliferating. Not all students can access such lessons due to their lack of socio-economic capital, therefore the practice of private tuition has widened the achievement gap among students. However, despite this, the assessment system is on the way to shifting from an exam-dominated system.

In current pre-service teacher education programs in Myanmar, the main assessment content is delivered in subjects on educational testing and measurement, compulsory for all students in teacher training universities. The content is normally related to the construction of the tests, for example, the functions of the tests and item analysis. Even though different forms of assessment—including formative assessment, performance assessment, and portfolio—are covered, the practical understanding and use of these assessments is still undeveloped. According to the findings of Hardman et al. (2016), teachers in Myanmar do not use A*f*L during teaching. For example, teachers do not use peer tutoring, and teachers do not seem to know how to build pupils' responses into subsequent questions. Therefore, this study aims to investigate the way in which PSTs can improve their assessment practices.

MATERIALS AND METHODS

Using a design-based research methodology, a needs-based professional development (PD) program for PSTs' assessment literacy was developed and delivered. Following the program, thirty PSTs in the intervention group were encouraged to implement the new A*f*L strategies during their practicum. Semi-structured individual interviews were undertaken with the

intervention group before and after their practicum in schools. The interviews were conducted to explore how PSTs applied their knowledge into their practice. For example, 'What assessment strategies have you tried out in class?' 'Why did you use ____ assessment strategy most frequently/least frequently? How did you use? Could you give me an example?'. In addition, lesson plans, observation checklists and audiotapes of their teaching for at least seven teaching periods were gathered from each PST during their practicum, so that they were able to reflect on their assessment practices with the help of these practicum data templates.

The needs-based PD program was grounded in a view of A*f*L literacy (Alonzo, 2016) that reflects the principles of A*f*L. The content of the PD program was adjusted based on the results of needs analysis that identified the current state of PST A*f*L literacy. The PD program includes four main parts: (i) A*f*L strategies; (ii) applying A*f*L to practice; (iii) developing teacher A*f*L literacy; and (iv) microteaching or peer-group practice teaching. This program was conducted over 2 months (a total of 36 h) with each session taking 2 h as presented in **Table 1**. Furthermore, the manner in which the program was provided was also an essential component of PST learning. Many courses in ITE are at odds with the underpinning principles of A*f*L (Timperley, 2014), however the present study followed Davison (2013) and Timperley (2014), ensuring the assessment program was grounded through an A*f*L approach. Thus, the workshop sessions in the program included initial 'sharing/reflection' to explore the background knowledge of the students and to encourage them to recall their previous experiences, and 'follow-up' to enable PSTs to reflect on what they had learned. All activities included in this program were based on the local context.

This study was conducted in one of the leading teacher training institutes in Myanmar. Fourth-year student teachers, who had already had experience of practice teaching in their third year, were chosen. A non-probability population sampling method was used due to the voluntary nature of participation. Before the data collection process, ethics approval was gained from the Institutional Review Board (IRB) of Ethics Committee and written permission was gained from the head of the participating university, Myanmar. Among thirty PSTs who expressed their interest to participate in this study, 10 PSTs (33%) were male and 20 PSTs (67%) were female. For their practicum teaching, 30 PSTs went to 17 practicum schools. They had varied total number of teaching period, one teaching period per day to more than three teaching periods per day which depends on the nature of their practicum school.

RESULTS

Following the strategies for qualitative data analysis described by Maxwell (2013), this paper presents the results of the thematic analysis of the semi-structured individual interviews before and after the practicum, with the data collected during the practicum used for triangulation. Five main themes emerged as enabling or constraining factors that influence PST assessment decision-making process: PST assessment knowledge, PST beliefs and values of using assessment, supervising teachers' influence, student responses and classroom realities. Grounded in a sociocultural approach to teacher agency (Priestley et al., 2015), these main themes were then classified into three dimensions of teacher agency: (1) the iterational dimension; (2) the projective dimension; and (3) the practical-evaluative dimension. In this study, the iterational dimension refers to the PSTs' assessment knowledge acquired through supplementary professional development in their ITE program. The projective dimension refers to the PSTs' aspirations for their profession and

TABLE 1 | Course content and structure of the Professional Development (PD) program.

Week	Content	Topic
Week 1	Part 1: A*f*L strategies	Session 1: understanding the interrelationship between assessment, teaching and learning
		Session 2: understanding assessment *for* learning (A*f*L)
Week 2		Session 3: framing learning intentions and success criteria
		Session 4: designing a rubric to improve student learning
Week 3		Session 5: involving learners in assessment (self- and peer-assessment)
		Session 6: giving effective feedback and feed-forward
Week 4		Session 7: using strategic questioning
		Session 8: using summative assessment in a formative way
Week 5	Part 2: application A*f*L to practice	Session 9: designing appropriate assessment strategies
		Session 10: planning learning and teaching experiences
Week 6		Session 11: enhancing the trustworthiness of an assessment
		Session 12: gathering assessment information
Week 7	Part 3: developing teacher A*f*L literacy	Session 13: evaluating and developing teacher assessment literacy
	Part 4: peer-group practice teaching	Session 14: peer-group practice teaching
Week 8		Session 15: peer-group practice teaching
		Session 16: peer-group practice teaching
Week 9		Session 17: peer-group practice teaching
		Session 18: peer-group practice teaching

for their students whilst the practical-evaluative dimension refers to the PSTs interactions with students, supervising teachers and classroom resources while on their final practicum.

Iterational Dimension: PST Assessment Knowledge

PST previous assessment knowledge is one of the key influences on PSTs' decision-making process. Based on their assessment knowledge gained through their professional learning, the PSTs prepared their A*f*L strategies and lesson plans. Some PSTs decided to use more assessment strategies to enhance students' learning. They adjusted their assessment strategies based on their knowledge of student backgrounds and learning needs. For example, PST 11 implemented learning intentions and success criteria, questioning strategies, feedback, self-assessment, and peer-assessment. She used flexible assessment activities and conducted the assessment taking into account the student's background. She put much effort into her preparation to use A*f*L strategies in her practicum:

> Before I give feedback to them, I have to know all the details. So, I have to prepare very well at night during the practicum. I have spent much time engaged in preparation. This makes me feel more confident in my teaching (PST 11, L 191–193).

Similarly, PST 10 prepared a detailed lesson plan of her assessment strategies, and thought she had been able to implement it effectively, taking into account possible student responses:

> From the beginning of preparing lesson plan, I pre think how I'm gonna teach and use assessment, so it is not much difficult. All lessons are taught in expected time range (PST 10, L 175,176).

Unlike PST 10, PST 27 did not prepare the lesson plan systematically to fit the duration of the teaching, for example, she did not set a time for each activity. Her teaching did not match with the lesson plan as she was uncertain when to finish the lessons:

> I aimed to teach as I intended in my lesson plan. But when I actually teach, I worry about not finishing all the lessons or having enough time and I didn't get to teach as I intended (PST 27, L 80–82).

As this was the second practicum for the PSTs, they compared their assessment practices with their first practicum experience. They highlighted how they had improved their use of assessment strategies in this second practicum as well as their assessment decision-making skills. For example, PST 8 commented on her improvement in setting success criteria and learning intention to improve student learning:

> Last time I also taught Myanmar subject (her first practicum), which needs much roles of teachers' explanation. This time I planned how I'm gonna use assessment, setting success criteria and learning intention before I get to teach. It's really effective for me letting me know the important facts (PST 8, L 207–212).

Similarly, PST 19 commented how she could better implement feedback in this second practicum. In her first practicum, she decided not to use feedback as she did not have enough assessment knowledge and skills. In this second practicum, she was satisfied with her use of feedback, and noted the progress of her use and her students' improvement in applying feedback in their learning:

> In the first practicum, I could not even assess their papers, not even got to the stage giving feedback. I was just lazy, think teaching was the main. But this time, I give feedback to let them know if they actually understand. I realize how to give feedback and note instantly (PST 19, L 222–228).

This section shows how PSTs' preparation in assessment before the practicum had an impact on PST successful implementation of A*f*L strategies. In addition, PST assessment knowledge helped them adjust implementation of assessment strategies based on student backgrounds and learning needs.

Projective Dimension: PST Beliefs and Values of Using Assessment

PST beliefs and values in using assessment is one of the key themes influencing PST classroom assessment decision-making. When PSTs had strong beliefs and values in relation to using assessment to improve students' learning, their positive efforts in using appropriate assessment strategies could be seen in their practicum. In the same way, PSTs did not put much effort into their classroom assessment practices when they did not really believe in the benefits of using assessment strategies to improve learning.

Some PST were well-prepared for their use of assessment strategies as they had strong beliefs and values of using these assessments. For example, PST 8 described the effectiveness of using assessment strategies. She articulated feedback in her practicum based on her students' needs. At the end, she was satisfied in her use of assessment and her decision-making:

> The best part is that when I give them feedback, I understand how to make it interesting even writing in red pen. Most students don't like red, but I use it with trendy style, so they love it. Even though they see comments in red on their papers, they read them interestingly. I feel quite satisfied to see that they never make those mistakes again and put much effort on it (PST 8, L 75–81).

In addition, she implemented questioning strategies successfully. She could build the students' answers into subsequent questions, and articulated her students' progress:

> I am well-pleased with the assessment, especially the strategic questioning. Depending on what students respond, I like that I could lead them to get the correct answer themselves (PST 8, L 143–145).

However, some PSTs received negative responses from students as they could not see the positive benefits of their assessments. Their PSTs did not use flexible teaching activities, develop an environment of trust nor build students' interest in learning. For example, PST 17 did not implement even one A*f*L strategy because he was not passionate about his practice teaching:

> It's just practicum so I didn't think I have much responsibility. As the students were not obedient so I didn't go against them. I had to teach for only 2 weeks, so I didn't scold them much and I wasn't too strict (PST 17, L 112–116).

Like PST 17, PST 3 could not implement at least one A*f*L strategy successfully although she tried. Then she decided not to use these assessment strategies. She did not develop an environment of trust, did not undertake assessment taking into account student background, and did not clarify or correct students' misconceptions. The evidence can be seen in the following extract:

> I told them to ask for help from their peers if they don't understand something. If not, they can ask to me (Interviewer: So, did they come and ask you?). Yes, they came and asked me. Then, I referred to another student who knows that answer. I couldn't do the detail explanation because of . . . (PST 3, L 83–85).

This shows that positive beliefs and values about using assessment generally led to successful assessment practices in practicum, and more negative attitudes led to an avoidance of the use A*f*L strategies. This result is consistent with that of Izci and Caliskan (2017) who suggested that the experience of successful assessment practices through the positive personal effort of PSTs leads to improving PSTs' conceptions of assessment. To this end, these results confirm the association between PST personal effort and their successful assessment practices.

Practical-Evaluative Dimension: Supervising Teachers' Influence

While the practicum is important to improve PST assessment practices in their teaching, not surprisingly supervising teachers were one of the main influences on PST decision-making regarding A*f*L strategies. In this study supervisors could be divided into controlling or supporting. With controlling supervising teachers, two sub-themes emerged: (i) control over instructional strategies of PST teaching; and (ii) control over the lessons/curriculum that PSTs need to teach. Regarding supporting supervising teachers, two sub-themes emerged: (i) academic/professional support through sharing lesson plans, giving constructive feedback and discussing PSTs' teaching; and (ii) autonomy, the freedom to develop teaching and assessment practices.

Controlling Effect of Supervising Teachers

This study looked closely at the influence of the personal attributes of supervising teachers on PSTs: their supporting and controlling effects. The study showed that supervising teachers helped or hindered PST implementation of assessment strategies. More controlling supervising teachers were associated with developing tensions and a poor relationship between the supervising teacher and PST. PSTs who had supervising teachers who were very controlling in relation to instructional strategies and the lessons/curriculum, commented that they had to change their assessment decisions and they adjusted their assessment strategies. They could not use assessment strategies according to their lesson plan.

When supervising teachers controlled their instructional strategies, PSTs were not allowed to use assessment-based activities. For example, PST 12 planned to use questioning strategies, feedback, self-assessment and peer-assessment over a range of activities. However, her supervising teachers persisted in controlling her teaching. She commented on how her supervising teacher influenced her teaching:

> My supervising teacher told me to teach what I need to teach, like focusing on lessons, not on any extra activities. And she is not observing my teaching from outside of the classroom, she is even sitting in the class with the students (most of her teaching periods) so I don't get any chance to let students do any activities. Also, the students around her didn't concentrate on my teaching (PST 12, L 153–155).

Subsequently, PST 12 revealed that she could not use most assessment strategies as she expected and planned. In the middle of her practicum, she decided not to implement assessment strategies because of the tension with her supervising teacher. She was not satisfied with her use of A*f*L strategies although she recognized the importance and effectiveness of using assessment after the program.

As a result of such control over their teaching, some PSTs were not motivated to use assessment strategies. They could not make choose to use trustworthy assessment strategies to improve students' learning. For example, PSTs 7 and 28 were hesitant to use assessment strategies as their supervising teachers gave critical feedback in front of their students to control their use of assessment activities. At the end of their practicum, they were unenthusiastic about their teaching and their use of assessment-based activities. The controlling effect of their supervising teachers can be seen in the following extract:

> Before even taking the class, I felt uncomfortable worrying that I might get scolded by my supervising teacher. I am not free to teach at all. I am not satisfied with my teaching as I don't have much preparation time and I don't get to use much assessments. While assessments are in advance, the class might get noisy, so I am concerned about what the other class teachers think. That's why I didn't use assessment frequently (PST 7, L 97–100).
>
> While I ask my students to participate in assessment activities, the teachers always shout and scold at us saying "Keep the voice down, it disturbs other classes." He does that every 2 days. So, I have to think twice before I do activities (PST 28, L 87–94).

In terms of the controlling effect of supervising teachers on lesson content, PSTs mentioned that their supervising teachers were very strict about finishing lessons. They commented that when their supervising teachers asked three or four times to complete lessons in the practicum, it was hard to apply A*f*L strategies to improve student learning. They commented that they were forced to focus on the completion of lessons rather than the use of A*f*L strategies because of the controlling effect of their supervising teachers. For example:

> Before I started taking a class, I aimed to teach effectively to make sure students understand, by applying proper assessment. But I was instructed to teach up to their [supervising teachers'] expected curriculum, so I had to rush and even took extra classes. My aimed assessment plan was ruined (PST 14, L 33–38).

Having negative experiences with supervising teachers also led to negative consequences for the PSTs' teaching practice. Some PSTs commented that they received critical comments on their teaching, and they developed bad relationships with their supervising teachers. For example, PST 3 commented that she felt disappointed and unmotivated in her teaching because of criticism from her supervising teacher:

> She said she could not teach again what I had taught, and the exam was coming at the end of the month, so students were gonna fail. That's what she said. And I even ask myself am I the reason why students gonna fail? (PST 3, L 186–189)

These findings reflect those of Smith (2010) who also found that disagreement between student teachers and supervising teachers had a negative effect. PSTs have more challenges when their supervising teachers are controlling their assessment practices (Cavanagh and Prescott, 2007). These results are consistent with the literature, indicating that supervising teachers have an influence not only on PSTs' teaching (Spooner-Lane et al., 2009; Smith, 2010; Izadinia, 2016; Livy et al., 2016) but also on their authentic assessment practices in the classroom (Graham, 2005; Volante and Fazio, 2007; Absolum et al., 2009; Eyers, 2014; Jiang, 2015).

Supporting Effect of Supervising Teachers

With supportive supervising teachers, two sub-themes emerged in relation to their behaviors: (i) the provision of more autonomy, which gave PSTs the freedom to develop their teaching and assessment practices; and (ii) academic/professional support through sharing lesson plans, giving constructive feedback and discussing PSTs' teaching. In particular, PSTs who gained greater autonomy during their practicum better understood assessment strategies and continuously applied the results of any assessment activity to identify room for improvement in both their teaching and students' learning.

If PSTs had supportive supervising teachers who provided autonomy in their teaching, they could then make trustworthy decisions in using assessment to enhance students' learning. For example, PST 11 commented that her supervising teacher did not tightly control her teaching and gave her freedom regarding the use of assessment strategies. Therefore, she was able to choose appropriate assessment strategies based on students' responses.

> My supervising teacher didn't control my instructional strategies and the lesson/curriculum that I need to teach. She explained what lessons I need to finish within these 2 weeks at the beginning of my practicum. She gave me the autonomy. She just came to observe my teaching twice for assessment purposes (PST 11, L 53–57).

A comparison of these findings with those of other studies (Weaver and Stanulis, 1996; Moody, 2009) confirm that autonomy can create the opportunity for PSTs to improve their teaching during practicum. Hence, this study seems to reinforce the literature which suggests that PSTs need to have sufficient autonomy to improve their teaching and assessment practices.

Regarding academic support from the supervising teacher, very few PST received academic support such as sharing lesson plans and giving constructive feedback on their teaching practices. PST 9 and 20 were an exception, receiving such support from their supervising teachers.

> She supported by providing me with materials. For example, notes of the lesson which is related to the lessons of the curriculum. She showed me how she did it (PST 9, L 48–49).
>
> I got the support from them, for example, their notes of the lesson. Then, she advised me how I can do the teaching (PST 20, L 73–74).

In contrast to these PSTs, most PST did not get any professional support from their supervising teachers, such as engaging in a discussion about their teaching, although, PSTs wanted to such support during practicum. For example, PST 7 expected emotional support from his supervising teachers such as friendly and helpful guidance:

> When I had my first teaching period, she didn't tell me how to teach with regard to the curriculum, nor did she discuss with me or even introduce me to the students. That's when I become inactive (PST 7, L 179–182).

The PSTs expected to receive such support, including engaging in discussion about their teaching. This finding reinforces studies (Cherian, 2007; Caires et al., 2012) which indicated PST need emotional and caring support from their supervising teachers. In addition, this finding confirms the results of previous studies (Richards and Crookes, 1988; Volante and Fazio, 2007; Spooner-Lane et al., 2009) that found insufficient support by supervising teachers in PST practicums. Nguyen (2016) suggested that many supervising teachers will support PSTs only when they have problems during the practicum. However, like other studies (Jiang, 2015), this study found that PSTs only successfully engaged in experimentation with assessment practices if they had the support of their supervising teachers, especially positive and frequent support. Therefore, supervising teachers need to be prepared to support PST A*f*L assessment practices.

As can be seen from these findings, supervising teachers play an important role in PST practicum. These results are consistent with the literature, indicating that supervising teachers have an influence not only on PST's teaching (Spooner-Lane et al., 2009; Smith, 2010; Izadinia, 2016; Livy et al., 2016) but also on their authentic assessment practices in the classroom (Graham, 2005; Volante and Fazio, 2007; Absolum et al., 2009; Eyers, 2014; Jiang, 2015). Hence, it is important for supervising teachers to have a positive influence on PST assessment.

However, in Myanmar, where this study took place, there is no proper mentoring program for supervising teachers about how to be a good mentor and how to help PSTs in their practicum. Hence, the findings of this study suggest that supervising teachers should be provided with guidelines on how to support PST, especially in relation to PSTs' assessment practices and their classroom assessment decision-making during the practicum.

Practical-Evaluative Dimension: Student Responses

This study also found that students' responses influenced PST classroom assessment decision-making. When PSTs

implemented A*f*L strategies in their practicum, they received various positive, negative or a combination of both responses from their students. PSTs made decisions to adjust their use of assessment strategies or to stop using them, based on students' responses. It is possible that students' responses depended on how students saw their PST and to what extent their PSTs were engaged, reflective, and how much effort and passion they put into their teaching.

When PSTs had positive responses from students, they decided to use assessment strategies frequently to improve their students' learning. Such PSTs mentioned their students' active participation in assessment activities, ongoing discussion about the lesson after the practicum and positive comments from their students. Some PSTs asked for feedback from their students after the practicum so they could reflect on their teaching. They felt satisfied about their decision to use A*f*L strategies if they received positive comments from students. For example:

> In their comments which are anonymous, they (her students) said they had a clear understanding after engaging in all assessment activities. If not, they could not decide the correct answer (PST 10, L 46–48).

Therefore, students' engagement and their progress were positive influencing factors which helped PSTs use appropriate assessment strategies. For example, PST 11 commented on her students' engagement in assessment practices. Although her students were not familiar with the strategies, the progress of her students could be seen through the outcome of using of them:

> Even if I forget to give feedback, they remind me to do it (PST 11, L 167–168).

> Some come along with questions saying that they think it ought to be another way. And I think this is kind of showing their engagement and you can see their interest (PST 11, L 71–73).

As a result, she decided to use these strategies till the end of the practicum. These results are consistent with those of Absolum et al. (2009) who noted the positive effect of active student-teacher collaboration in assessment practices. However, some PSTs decided not to use these assessments when they had unexpected challenges from their students. For example, some students did not want to give feedback to their peers. In this case, their students gave feedback to PSTs. For example:

> After they have done the peer-assessment, they never wanted to give feedback to other students. They always come to me, show me what they've done and tell me how they think. That's not what I expected them to do in their peer-assessment (PST 2, L 64–67).
>
> They don't wanna give feedback to their peers. What they worry is about that they might assess others wrongly. They never write negative feedback although it is wrong. They worry that they might annoy other students. Maybe because they have never done that before (PST 1, L 49–53).

However, some students had arguments with their peers based on the feedback. Consequently, some PSTs commented that using peer-assessment did not work well according to their lesson plan. They stopped using peer-assessment as they could not control the classroom situation:

> I thought they would love to undertake peer-assessment before practicum. When I actually do it, they argue a lot. Therefore, I think before I actually assign peer-assessment to them. I should probably change their attitude first (PST 6, L 115–117).

Therefore, this study showed that students' responses toward A*f*L strategies influenced the success or failure of their assessment practices. This result confirms the results of previous studies (Elwood and Klenowski, 2002; Absolum et al., 2009; Jiang, 2015) which found that the influence of students on PST assessment practices was fundamental for effective A*f*L. Charteris and Dargusch (2018) also observed that students are crucial in shaping and reshaping PST assessment practices during the practicum.

During classroom interactions, the way students responded to their teachers was related to how teachers treated them in terms of using assessment. It follows then that students responded positively to PSTs in terms of the overall A*f*L strategies, and each A*f*L strategy when they saw positive efforts from their PSTs. On the other hand, students responded negatively when they saw the negative efforts of their PSTs. The performance of PSTs is one of the causes of positive and negative student responses. However, it should be noted that this study focuses on the results from the perspectives of PSTs.

Practical-Evaluative Dimension: Classroom Realities

During PST assessment practices, classroom realities emerged as one of the influencing factors on PST assessment decision-making. The classroom setting of the school, the number of teaching periods, and the time of day of the particular period influenced assessment decision-making during the practicum.

The classroom setting of the school was one of the influential factors in PSTs' assessment practices. Unless they had enough space, PSTs could not implement the assessment-based activities effectively. Many PSTs needed to group students or rearrange students in a lecture-oriented classroom when they implemented assessment-based activities. For this reason, some PSTs commented that they were negatively influenced by the classroom setting. They stopped using self- and peer-assessments because of the negative influence of space in the classroom. For example:

> The classroom isn't wide enough. It can't rearrange desks and chairs for activities. It's just wasting time (PST 28, L 29–30).
>
> The classroom is not large enough for a teacher to walk through (PST 14, L 20).

In contrast, some PSTs reported that they had enough space to implement assessment-based activities. For example, PST 11 commented that her classroom was wide enough to implement most A*f*L strategies:

> The classroom has enough space for 36 students to perform assessment-based activities, so it doesn't matter to be noisy (PST 11, L 47–49).

The data suggests that PSTs need enough space in their classroom to do key assessment activities. Therefore, the physical

setting of the school help PSTs choose appropriate assessment strategies based on students' needs.

A second classroom reality was the number of teaching periods, which the PSTs could not control. This study found a variety of PSTs' experience regarding the number of teaching periods. Some PSTs had more than three teaching periods per day while some had less than one teaching period per day. When PSTs had fewer teaching periods in their practicum, they did not get the chance to implement more assessment strategies or assessment practices. For example, PST 4 mentioned that:

> I got just five teaching periods for the whole practice teaching. This is not enough to get an experience of assessment practices (PST 4, L 222,223).

This study shows that PST can experiment with more A*f*L strategies when they have more teaching periods where they have a chance to practice their assessment knowledge and skills. These results are consistent with those of Mitton-Kukner and Orr (2014) who found that the length of the teaching period is one of the influences on the PST practicum.

In terms of time of day, four PSTs commented that they have positive and negative influences in their assessment practices, for example, earlier and later time of day. They commented that the time of day had an effect on students' involvement in assessment practices. For example:

> Most of the students could not concentrate on lessons at the last period of a day. During last class, they just wanna finish the class. Only the students sitting on the front seats pay attention on teaching (PST 1, L 119–121).
>
> As my class time is second period so that's ok. Once one physics teacher requested me to switch my class with hers because she wanted her students to learn in fresh minds. And I ended up taking afternoon class which is period after lunch break. I could not teach properly on that days. I had to wait may be 15 min because there was complete chaos when I entered the classroom (PST 8, L 30–35).

These PSTs commented that if they had the earlier time of day, they could decide to implement more A*f*L strategies and their students could engage more in assessment practices. If PSTs had teaching periods later in the day, students did not actively engage in assessment activities. Therefore, an earlier time of day had a positive influence on students' responses in PSTs' assessment practices while a later time of day had a negative influence on students' responses. This suggests that time of the day is one of the classroom realities that PST could not control. Therefore, PST should be equipped with many opportunities to experiment with assessment practices in the practicum. In addition, there is no literature on the association between these factors of classroom realities and PST assessment practices. Therefore, further studies need to be undertaken which can take these influences into account.

DISCUSSION AND CONCLUSION

Drawing on the nature of teacher agency, this study has enabled us to understand the factors which can influence PSTs' assessment decision-making process and the extent to which PSTs can exercise agency by engaging with the influencing factors of school context, available resources, and their beliefs and experiences. The factors which influenced classroom assessment decision-making found in this study were (i) the iterational dimension: PST assessment knowledge (ii) the projective dimension: PST beliefs and values of using assessment, and (iii) the practical evaluative dimension: their supervising teachers, students' responses, and classroom realities. These influences on teacher assessment decision-making are somewhat aligned with previous studies (McMillan and Nash, 2000; McMillan, 2003) which also demonstrated the influence of external factors, including state accountability testing, district policies, and parents. However, as the role of PSTs in the practicum does not include working with parents and school leaders within such a short period of practice teaching in Myanmar, the influence of parents and district policies was not explored.

In decision-making processes, there is also tension between PSTs' beliefs and values and external influences: stakeholder, student responses, and classroom realities which is consistent with the results of Black and Wiliam (1998a) and McMillan (2003). PSTs who have a good sense of how A*f*L operates, in using on-going assessment and using the results to make decisions including adjustment of learning and teaching activities, developed tension with supervising teachers who exerted strong control over their practicum. In addition, when PSTs are negatively influenced by one or more of these factors, they could not make appropriate assessment decisions to improve students' learning. Those PSTs who gained greater autonomy during their practicum better understood assessment strategies and continuously applied the results of any assessment activity to identify room for improvement both in their teaching and students' learning. Therefore, this study contributes to recent literature on teacher agency (Priestley et al., 2013, 2015; Buchanan, 2015; Loutzenhesier and Heer, 2017) which has argued that teacher agency during PST assessment practices is heavily impacted by particular contextual factors. Teacher agency has emerged through their engagement with the environment which is consistent with previous studies (Biesta and Tedder, 2007; Priestley et al., 2013; van der Nest et al., 2018). Teachers exercised their attributes in engaging with that specific context, for example, modifying the assessment strategies. PSTs responded differently to these influences in accordance with the findings of Verberg et al. (2016).

In general, teacher training institutes or colleges need to understand the essential role of authentic assessment practices in classrooms for PSTs. Participating in assessment practices develops a sense of agency that they can engage with real students in classroom. The findings of this study show that PSTs improved their classroom assessment decision-making through working with students. Therefore, teacher training institutes or colleges need to ensure that student teachers have an opportunity to practice and reflect on their assessment during practicum. To improve PST classroom assessment decision-making in their assessment practices, cooperation between teacher training institutes and school practicum schools must be improved. It is important that PSTs, teacher educators

and the other key stakeholders from the practicum school can speak the same language, especially in PST teaching where influences on PST assessment practices are interactive. However, in Myanmar, there is less contact between teacher educators and the practicum school in terms of improving PST assessment practices and teaching than in assessing PST teaching generally. This suggests that teacher educators, supervising teachers and PSTs should cooperate more at the beginning of the practicum. This study provides a better understanding of how to improve PST assessment decision-making in their assessment practices through addressing the interactive nature of assessment influences.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the HREA Panel B: Arts, Humanities & Law of UNSW Australia and Yangon University of Education, Myanmar. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

CO: conceptualization, Professional Development (PD) program design and implementation, PD program delivery, methodology, data collection, transcription, data analysis and interpretation, writing-reviewing, and editing. DA and CD: conceptualization, PD program design and implementation, methodology, data analysis and interpretation, writing-reviewing, and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by a Presidential Scholarship, Myanmar.

REFERENCES

Absolum, M., Flockton, L., Hattie, J., Hipkins, R., and Reid, I. (2009). *Directions for Assessment in New Zealand: Developing Students' Assessment Capabilities*. Wellington: Ministry of Education.

Alonzo, D. (2016). *Development and Application of a Teacher Assessment for Learning (AfL) Literacy Tool*. Doctoral dissertation, The University of New South Wales, Kensington, NSW.

Assessment Reform Group (2002). *Assessment for Learning: 10 Principles*. Available online at: www.assessment-reform-group.org.uk (accessed August 2, 2016).

Aung, W., Hardman, F., and Myint, A. A. (2013). *Development of a Teacher Education Strategy Framework Linked to Pre- and in-Service Teacher Training in Myanmar*. San Diego, CA: The Institute For Effective Education.

Biesta, G., and Tedder, M. (2007). Agency and learning in the lifecourse: towards an ecological perspective. *Stud. Educ. Adults* 39, 132–149. doi: 10.1080/02660830.2007.11661545

Billett, S. (2004). Workplace participatory practices. *J. Workplace Learn.* 16, 312–324. doi: 10.1108/13665620410550295

Black, P., and Wiliam, D. (1998a). Assessment and classroom learning. *Assess. Educ.* 5, 7–74. doi: 10.4324/9781315123127-3

Black, P., and Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan* 80, 139–148. doi: 10.1002/hrm

BOSTES (2016). *Learning Assessment: A Report on Teaching Assessment in Initial Teacher Education in NSW*. Sydney, NSW: NSW Government.

Bowers, A. J. (2009). Reconsidering grades as data for decision making: more than just academic knowledge. *J. Educ. Administr.* 47, 609–629. doi: 10.1108/09578230910981080

Brookhart, S. M. (1991). Grading practices and validity. *Educ. Meas. Issues Pract.* 10, 35–36. doi: 10.1111/j.1745-3992.1991.tb00182.x

Buchanan, R. (2015). Teacher identity and agency in an era of accountability. *Teach. Teach.* 21, 700–719. doi: 10.1080/13540602.2015.1044329

Caires, S., Almeida, L., and Vieira, D. (2012). Becoming a teacher: Student teachers' experiences and perceptions about teaching practice. *Eur. J. Teach. Educ.* 35, 163–178. doi: 10.1080/02619768.2011.643395

Campbell, C. (2013). "Research on teacher competence in classroom assessment," in *SAGE Handbook of Research on Classroom Assessment*, ed. J. H. McMillan (Thousand Oaks, CA: SAGE).

Cavanagh, M., and Prescott, A. (2007). "Professional experience in learning to teach secondary mathematics: incorporating pre-service teachers into a community of practice," in *Proceedings of the 30th Annual Conference of the Mathematics Education Research Group of Australasia*, Vol. 1, eds J. Watson and K. Beswick (Adelaide: MERGA), 182–191.

Charteris, J., and Dargusch, J. (2018). The tensions of preparing pre-service teachers to be assessment capable and profession-ready. *Asia Pac. J. Teach. Educ.* 46, 354–368. doi: 10.1080/1359866X.2018.1469114

Cheng, L., and Sun, Y. (2015). Teachers grading decision making: multiple influencing factors and methods. *Lang. Assess. Q.* 12, 213–233. doi: 10.1080/15434303.2015.1010726

Cherian, F. (2007). Learning to teach: teacher caniddates reflect on the relational, conceptual, and contextual influences of responsive mentorship. *Can. J. Educ.* 30, 25–46. doi: 10.2307/20466624

Cramer, E. D., Little, M. E., and McHatton, P. A. (2014). Demystifying the data-based decision-making process. *Action Teach. Educ.* 36, 389–400. doi: 10.1080/01626620.2014.977690

Davison, C. (2013). "Innovation in assessment: common misconceptions and problems," in *Innovation and Change in English Language Education*, eds K. Hyland and L. L. Wong (Oxon: Routledge), 263–275.

Davison, C. (2015). "Enhancing teacher assessment literacy: practising what we preach," in *Paper Presented at 2015 Assessment in Schools Conference*, (Sydney, NSW).

Davison, C. (2019). "Using assessment to enhance learning in English language education," in *Second Handbook of English Language Teaching, Springer International Handbooks of Education*, ed. X. Gao (Cham: Springer Nature Switzerland), 433–454. doi: 10.1007/978-3-030-02899-2_21

DeLuca, C., and Volante, L. (2016). Assessment for learning in teacher education programs: navigating the juxtaposition of theory and praxis. *J. Int. Soc. Teach. Educ.* 20, 1–13. doi: 10.1016/j.tate.2013.02.001

Elwood, J., and Klenowski, V. (2002). Creating communities of shared practice: the challenges of assessment use in learning and teaching. *Assess. Evaluat. High. Educ.* 27, 243–256. doi: 10.1080/02602930220138606

Eyers, G. (2014). *Preservice Teachers' Assessment Learning: Change, Development and Growth*. Doctoral dissertation, The University of Auckland, Auckland.

Graham, P. (2005). Classroom-based assessment: changing knowledge and practice through preservice teacher education. *Teach. Teach. Educ.* 21, 607–621. doi: 10.1016/j.tate.2005.05.001

Grainger, P., and Adie, L. (2014). How do preservice teacher education students move from novice to expert assessors? *Aust. J. Teach. Educ.* 39, 1–18. doi: 10.14221/ajte.2014v39n7.9

Greenberg, J., and Walsh, K. (2012). *What Teacher Preparation Programs Teach About K – 12 Assessment: A Review*. Washington, DC: National Council on Teacher Quality.

Hardman, F., Stoff, C., Aung, W., and Elliott, L. (2016). Developing pedagogical practices in Myanmar primary schools: possibilities and constraints. *Asia Pac. J. Educ.* 36, 98–118. doi: 10.1080/02188791.2014.906387

Hattie, J. (2008). *Visible Learning: A Synthesis of Over 800 Meta-analysis Relating to Achievement*. Milton Park: Routledge.

Heck, D. (2020). "Talking about rubrics in higher education: exploring academic agency," in *Facilitating Student Learning and Engagement in Higher Education through Assessment Rubrics*, eds P. Grainger and K. Weir (Newcastle upon Tyne: Cambridge Scholars Publishing).

Heck, D., Willis, A., Simon, S., Grainger, P., and Smith, K. (2020). "Becoming a teacher: scaffolding post-practicum reflection," in *Enriching Higher Education Students' Learning through Post-work Placement Interventions, Professional and Practice-based Learning*, eds S. Billett, J. Orrell, D. Jackson, and F. Valencise-Forrester (Cham: Springer), 173–188. doi: 10.1007/978-3-030-48062-2

Hill, M. F., and Eyers, G. (2016). "Moving from student to teacher," in *Handbook of Human and Social Conditions in Assessment*, eds G. T. L. Brown and L. R. Harris (Milton Park: Routledge).

Hill, M. F., Smith, L. F., Cowie, B., Gilmore, A., and Gunn, A. (2013). *Preparing Initial Primary and Early Childhood Teacher Education Students to Use Assessment: Final Summary Report*. Wellington: Teaching and Learning Research Initiative.

Izadinia, M. (2016). Student teachers' and mentor teachers' perceptions and expectations of a mentoring relationship: do they match or clash? *Prof. Dev. Educ.* 42, 387–402. doi: 10.1080/19415257.2014.994136

Izci, K., and Caliskan, G. (2017). Development of prospective teachers' conceptions of assessment and choices of assessment tasks. *Int. J. Res. Educ. Sci.* 3, 464–474. doi: 10.21890/ijres.327906

Jiang, H. (2015). *Learning to Teach with Assessment: A Student Teaching Experience in China*. Cham: Springer.

Kippers, W. B., Wolterinck, C. H. D., Schildkamp, K., Poortman, C. L., and Visscher, A. J. (2018). Teachers' views on the use of assessment for learning and data-based decision making in classroom practice. *Teach. Teach. Educ.* 75, 199–213. doi: 10.1016/j.tate.2018.06.015

Klenowski, V. (2013). Towards improving public understanding of judgement practice in standards-referenced assessment: an Australian perspective. *Oxford Rev. Educ.* 39, 36–51. doi: 10.1080/03054985.2013.764759

Livy, S. L., Vale, C., and Herbert, S. (2016). Developing primary pre-service teachers' mathematical content knowledge during practicum teaching. *Aust. J. Teach. Educ.* 41, 1–23. doi: 10.1007/s13394-018-0252-8

Loutzenhesier, L., and Heer, K. (2017). "Unsettling habitual ways of teacher education through 'post-theories' of teacher agency," in *The SAGE Handbook of Research in Teacher Education*, eds D. Clandinin and J. Husu (Thousand Oaks, CA: SAGE Publication), 317–331. doi: 10.4135/9781526402042.n18

Maber, E. J. T., Oo, H. W. M., and Higgins, S. (2018). "Understanding the changing roles of teachers in transitional Myanmar," in *Sustainable Peacebuilding and Social Justice in Times of Transition*, eds M. Lopes Cardozo and E. Maber (Springer), 117–139. doi: 10.1007/978-3-319-93812-7_6

Maxwell, J. A. (2013). *Qualitative Research Design: An Interactive Approach*, 3rd Edn. Thousand Oaks, CA: SAGE.

Mccall, V. A. (2018). *The Decision-Making Process Used to Determine Formative Assessment Strategies and Subsequent Instructional Design*. Aurora, IL: Aurora University.

McGee, J., and Colby, S. (2014). Impact of an assessment course on teacher candidates' assessment literacy. *Action Teach. Educ.* 36, 522–532. doi: 10.1080/01626620.2014.977753

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educ. Meas. Issues Pract.* 20, 20–32. doi: 10.1111/j.1745-3992.2001.tb00055.x

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: implications for theory and practice. *Educ. Meas. Issues Pract.* 22, 34–43. doi: 10.1111/j.1745-3992.2003.tb00142.x

McMillan, J. H., and Nash, S. (2000). "Teacher classroom assessment and grading practices decision making," in *Paper Presented at the Annual Meeting of the National Council on Measurement in Education*, (New Orleans, LA), 39.

Metro, R. (2015). "Students and teachers as a agents of democratization and national reconciliation in Burma," in *Contemporary Burma/ Myanmar*, eds R. Egreteau and F. Robinne (Singapore: NUS Press), 209–223. doi: 10.2307/j.ctv1ntgbt.15

Mitton-Kukner, J., and Orr, A. M. (2014). Making the invisible of learning visible: pre-service teachers identify connections between the use of literacy strategies and their content area assessment practices. *Alberta J. Educ. Res.* 60, 403–419.

Moody, J. (2009). Key elements in a positive practicum: insights from Australian post-primary pre-service teachers. *Irish Educ. Stud.* 28, 155–175. doi: 10.1080/03323310902884219

Nguyen, L. T. H. (2016). *Development of Vietnamese Pre-service EFL Teachers' Assessment Literacy*. Doctoral thesis, Victoria University of Wellington, Wellington.

Ogan-Bekiroglu, F., and Suzuk, E. (2014). Pre-service teachers' assessment literacy and its implementation into practice. *Curriculum J.* 25, 344–371. doi: 10.1080/09585176.2014.899916

Piro, J. S., Dunlap, K., and Shutt, T. (2014). A collaborative data chat: teaching summative assessment data use in pre-service teacher education. *Cogent Educ.* 1, 1–24. doi: 10.1080/2331186X.2014.968409

Piro, J. S., and Hutchinson, C. J. (2014). Using a data chat to teach instructional interventions: student perceptions of data literacy in an assessment course. *New Educ.* 10, 95–111. doi: 10.1080/1547688X.2014.898479

Popham, W. J. (2011). Assessment literacy overlooked: a teacher educator's confession. *Teach. Educ.* 46, 265–273. doi: 10.1080/00405840802577536

Priestley, M., Biesta, G., and Robinson, S. (2013). "Teachers as agents of change: teacher agency and emerging models of curriculum," in *Reinventing the Curriculum: New Trends in Curriculum Policy and Practice*, eds M. Priestley and G. Biesta (London: A&C Black).

Priestley, M., Biesta, G., and Robinson, S. (2015). *Teacher Agency: An Ecological Approach*. London: Bloomsbury.

Reeves, T. D., and Chiang, J. L. (2018). Online interventions to promote teacher data-driven decision making: Optimizing design to maximize impact. *Stud. Educ. Evaluat.* 59, 256–269. doi: 10.1016/j.stueduc.2018.09.006

Reeves, T. D., and Honig, S. L. (2015). A classroom data literacy intervention for pre-service teachers. *Teach. Teach. Educ.* 50, 90–101. doi: 10.1016/j.tate.2015.05.007

Richards, J. C., and Crookes, G. (1988). The practicum in TESOL. *TESOL Q.* 9, 9–27. doi: 10.2307/3587059

Siegel, M. A., and Wissehr, C. (2011). Preparing for the plunge: preservice teachers' assessment literacy. *J. Sci. Teach. Educ.* 22, 371–391. doi: 10.1007/s10972-011-9231-6

Smith, K. (2010). Assessing the Practicum in teacher education – do we want candidates and mentors to agree? *Stud. Educ. Evaluat.* 36, 36–41. doi: 10.1016/j.stueduc.2010.08.001

Spooner-Lane, R., Tangen, D., and Campbell, M. (2009). The complexities of supporting Asian international pre-service teachers as they undertake practicum. *Asia Pac. J. Teach. Educ.* 37, 79–94. doi: 10.1080/13598660802530776

Stritikus, T. T. (2003). The interrelationship of beliefs, context, and learning: the case of a teacher reacting to language policy. *J. Lang. Ident. Educ.* 2, 29–52. doi: 10.1207/S15327701JLIE0201_2

Timperley, H. (2014). "Using assessment information for professional learning," in *Designing Assessment for Quality Learning*, eds C. Wyatt-Smith, V. Klenowski, and P. Colbert (Dordrecht: Springer), 137–149. doi: 10.1007/978-94-007-5902-2_9

Tin, H. (2000). Myanmar education: status, issues and challenges. *J. Southeast Asian Educ.* 1, 134–162.

van der Nest, A., Long, C., and Engelbrecht, J. (2018). The impact of formative assessment activities on the development of teacher agency in mathematics teachers. *S. Afr. J. Educ.* 38, 1–10. doi: 10.15700/saje.v38n1a1382

van Phung, D. (2018). *Variability in Teacher oral English Language Assessment*. Sydney, NSW: The University of New South Wales.

Verberg, C. P. M., Tigelaar, D. E. H., van Veen, K., and Verloop, N. (2016). Teacher agency within the context of formative teacher assessment: an in-depth analysis. *Educ. Stud.* 42, 534–552. doi: 10.1080/03055698.2016.1231060

Vogt, K., and Tsagari, D. (2014). Assessment literacy of foreign language teachers: findings of a European study. *Lang. Assess. Q.* 11, 374–402. doi: 10.1080/15434303.2014.960046

Volante, L., and Fazio, X. (2007). Exploring teacher candidates' assessment literacy: Implications for teacher education reform and professional development. *Can. J. Educ.* 30, 749–770. doi: 10.2307/20466661

Weaver, D., and Stanulis, R. N. (1996). Negotiating preparation and practice: student teaching in the middle. *J. Teach. Educ.* 47, 27–36. doi: 10.1177/0022487196047001006

Willis, J. (2007). Assessment for learning – why the theory needs the practice. *Int. J. Pedagogies Learn.* 3, 52–59. doi: 10.5172/ijpl.3.2.52

Xu, Y., and Brown, G. T. L. (2016). Teacher assessment literacy in practice: a reconceptualization. *Teach. Teach. Educ.* 58, 149–162. doi: 10.1016/j.tate.2016.05.010

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

frontiers in Education

ORIGINAL RESEARCH
published: 07 April 2021
doi: 10.3389/feduc.2021.631185

Assessment Conceptions and Practices: Perspectives of Primary School Teachers and Students

*Vera Monteiro *, Lourdes Mata and Natalie Nóbrega Santos*

Centro de Investigação em Educação, ISPA – Instituto Universitário, Lisbon, Portugal

Students' and teachers' conceptions of assessment are important because they guide how teachers' assessments are implemented in the classroom and determine how students study. This multiple-case design study examined 1) how teachers and students view assessment, 2) how teachers assess their students' learning, and 3) the similarities and disparities that occur when students' and teachers' conceptions and teachers' practices of assessment are compared. Data were obtained from five third grade classes, involving a total of five teachers and 82 students. Data were gathered through individual interviews with teachers and focus group discussions with students. Classroom observations and documents produced by the students (worksheets and tests) during maths lessons were also analyzed. The results of the content analysis of the data indicate that teachers mostly conceive assessment as being for improvement, while their assessment practices and students' conceptions focus on school and student accountability. The results obtained lead us to suggest that students' conceptions of assessment are constructed from their classroom assessment experiences. The study also suggests that teachers adopt conceptions of assessment inconsistent with their practices, that allow them to work within social and contextual constraints.

Keywords: assessment practices, primary school, teacher, students, assessment conceptions

OPEN ACCESS

Edited by:
Chris Davison,
University of New South Wales, Australia

Reviewed by:
Ibrahim Burak Ölmez,
University of Southern California, Los Angeles, United States
Peter Ralph Grainger,
University of the Sunshine Coast, Australia

****Correspondence:***
Vera Monteiro
veram@ispa.pt

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

Received: *19 November 2020*
Accepted: *09 February 2021*
Published: *07 April 2021*

Citation:
Monteiro V, Mata L and Santos NN (2021) Assessment Conceptions and Practices: Perspectives of Primary School Teachers and Students. Front. Educ. 6:631185. doi: 10.3389/feduc.2021.631185

INTRODUCTION

Classroom assessment has been a topic of interest for researchers in recent years. Focusing on assessment is important for the development of teaching and learning processes. Assessment enables teachers and students to draw inferences from the information obtained and act accordingly. Such actions may aid in making the necessary improvements to teaching and learning, or simply provide a picture in time of students' competence or achievement (Black and Wiliam, 2018).

The study of teachers' and students' conceptions of assessment is an important topic within the domain of assessment research. According to Brown (2008, p. 9), "conceptions of assessment refer to the perceptions people have about assessment, based on their experiences with and of assessment." Teachers' conceptions of assessment are significant because clear evidence exists that these beliefs strongly influence how teachers assess their students' learning and achievements (Vandeyar and Killen, 2007; Brown, 2008; Brown et al., 2009b; Opre, 2015). In addition, conceptions can also influence their classroom practices, such as instructional techniques and motivational strategies (Barnes et al., 2017). Students' conceptions of assessment are also important, since it is known that their beliefs guide and determine how they study (Brown and Hirschfeld, 2007; Brown and Harris, 2012).

Though this area of research has wide-ranging implications for the teaching and learning process, little is known about the conceptions of students and teachers in primary school, and how these conceptions are related to teachers' assessment practices. Therefore, the primary objective of the present study was to investigate whether primary school students' and teachers' conceptions are aligned with teachers' practices, and to discuss the implications for teaching and learning of this alignment, or its absence.

Teachers' Conceptions of Assessment

According to Brown (2004, p. 303), "all pedagogical acts, including teachers' perceptions and evaluations of student behavior and performance (i.e., assessment), are affected by the conceptions teachers have about many educational artifacts, such as teaching, learning, assessment, curriculum, and teacher efficacy." It is important to analyze this relationship when teachers' conceptions need to be changed, as in the case of reformulations in a country's education system with consequences for the student assessment system.

In his multiple studies, Brown (2004, 2008) found that teachers conceive assessment as having four major purposes. The first conception relies on the idea that assessment improves both teaching and students' learning. Hence, assessment should provide effective feedback, be enjoyable, be felt as something positive that helps students improve, and be inclusive and integrated with the teaching and learning process. A second conception views assessment as making students accountable through scoring, grading, or certification. This means that assessment is used to categorize, differentiate, make social comparisons, and determine whether students have met standards. A third conception views assessment as making schools and teachers accountable, and therefore providing information about the quality of education. The fourth conception relies on the belief that assessment is irrelevant. Here, assessment is seen as inaccurate and bad for students, and is ignored by teachers. In line with this definition of assessment conceptions, Brown (2008) constructed the Teachers' Conceptions of Assessment questionnaire (TCoA).

Research using the TCoA with New Zealand and Queensland primary teachers showed that teachers mostly agreed that assessment improved teaching and learning but disagreed that assessment was for student accountability. They also rejected the conception that assessment was irrelevant (Brown, 2008).

Another approach has been proposed by Remesal (2011), who sees teachers' assessment conceptions as a combination of four aspects: assessment effects on teaching, on learning, on students' certification of learning, and on teachers' accountability. According to the author, assessment can be viewed as being on a continuum with a formative-regulatory pole (pedagogical) and a non-regulatory social pole (societal), and two or three mixed conceptions in between (Brown and Remesal, 2017). When comparing primary and secondary teachers, Remesal (2009, 2011) found that the pedagogical conception of assessment (extreme and mixed forms) predominated among primary education teachers, whereas the accounting conception (societal and accrediting conceptions—extreme and mixed forms) predominated among secondary teachers' conceptions. The author hypothesized that these conceptions could be related to the structure of the educational system and external assessment policy demands in Spain.

Azis (2015) proposed an approach in which conceptions of assessment can be distributed on a continuum of different purposes. At one end of the continuum is Assessment for Learning (AfL), also called formative assessment (Brown and Remesal, 2017), or the pedagogical pole (Remesal, 2007). Here, assessment is aimed at promoting students' learning and providing teachers and students with the information needed to modify teaching and learning strategies (Black and Wiliam, 2018). At the other end of the continuum is Assessment of Learning (AoL), also called summative assessment (Brown and Remesal, 2017), or the societal pole (Remesal, 2007). Here, the focus is on high-stakes accountability, ranking, grading, and/or certification. Between these poles, we find mixed conceptions of the purposes of assessment (Azis, 2015). This approach has some similarities with those of Brown (2004, 2008; Harris and Brown, 2009) and Remesal (2006, 2011). In his article about teachers' conceptions of assessment, Azis (2012) reviewed numerous studies on this subject conducted in six different countries. The results revealed that all teachers believed that assessment improves learning and that assessment relates to school accountability. The author suggested that the six different countries in the review interpret improvement in different ways, being determined by factors such as curriculum level, government policy on education and the experience of teachers.

Students' Conceptions of Assessment

Much of the research on students' conceptions of assessment has also been conducted by Brown and his colleagues (e.g., Brown and Hirschfeld, 2007; Brown et al., 2009a; Brown and Harris, 2012), primarily with secondary and university students. References to such research with primary school pupils are scarce. This gap in the literature needs to be closed, because what students think about assessment mediates their learning and achievement and has consequences for how they participate in assessment tasks.

A review of the literature on students' conceptions of assessment in general (Brown and Hirschfeld, 2007; Brown 2008; Brown et al., 2009a; Brown and Harris, 2012) identified four different purposes for assessment: 1) improvement: assessment led to improvements in learning and teaching; 2) external attribution: assessment is linked to external attributes of the student, such as their future performance or job, their intelligence, and the quality of the school they attend; 3) affect: assessment has a positive emotional impact on students; and 4) irrelevance: assessment is oppressive, inaccurate, and ignored by students.

Remesal (2006, 2009) is one of the few authors to have studied this topic among primary pupils. Based on the categories used to study elementary teachers' assessment beliefs (referred to in the previous section), Remesal (2006) defined three categories of students' conceptions: in the first, students assigned a predominantly regulatory function to assessment; in the second, the predominant function was certification; and in the

third, the students did not assign any function. The results point to a balancing of the two main assessment functions (regulatory, 44.4% agreed; certification, 41.7% agreed), and predominant disagreement with the claim that assessment was irrelevant. Similar results were found with Finnish primary school students (Ämmälä and Kyrö-Ämmälä, 2018). It is apparent from these results that students have multidimensional conceptions of assessment and are aware of them from primary school onwards (Remesal, 2009).

Comparing Teachers' and Students' Conceptions of Assessment

What students and teachers think and believe about assessment is crucial for the efficiency of the teaching and learning process as well as for a shared understanding of the purposes of assessment in meeting learning and teaching goals. As Andersson (2016) observed, a shared understanding of what is being learning is essential if teachers are expected to help a student learn from their teaching experiences. In teacher-pupil interactions and in peer interactions, knowledge acquisition is dependent on the shared representation that both participants construct of the task and the context in which they are learning. According to Andersson (2016) and Gipps (1999), assessment can be seen as an intersubjectivity setting, where shared understanding between teacher and student is central to learning outcomes. Carless (2009) states that such shared understanding improves the assessment integrity and the quality of the student learning experience. It is therefore important that pupil and teacher conceptions of assessment are aligned.

Few studies have aimed to compare student and teacher conceptions of assessment (e.g., Remesal, 2006; Brown, 2008; Fletcher et al., 2012). Furthermore, these were mostly carried out with secondary or university students.

Remesal (2006) conducted one pioneering study with pupils and their primary school teachers and found differences in their conceptions. Teachers attributed to assessment a function closer to the pedagogical pole, while students presented a more balanced conception of assessment, attributing similar importance to a pedagogical conception and a societal one. Nevertheless, in most cases when pedagogical assessment was mentioned, teachers considered that it serves mostly for teaching improvement, since they believed that students are incapable of participating in the assessment process. The pupils, on the other hand, were of the view that assessment serves to improve not only the teaching but also their learning. They felt it helped them see what they have learned, whether they should try harder, and what they have not understood. According to Remesal (2006), students' and teachers' conceptions are more aligned when they present a pedagogical assessment conception and when teachers make the assessment criteria explicit. The author found more discrepancies when the assessment criteria were not explicit, regardless of teachers' assessment conceptions. Therefore, it seems that the degree to which the assessment criteria are made explicit exerts more influence on students' assessment conceptions than on teachers' conceptions.

The few researchers who have compared the assessment conceptions of teachers and their students have found that, in general, they differ. While students have a clear conception that assessment has a fundamental purpose—the certification of student learning, teachers' conceptions of assessment are not very clear but show a strong tendency toward the purpose of improving teaching and learning. Since teachers and students are directly involved in the same pedagogical process (assessment), it is strange that they perceive it to have different purposes. In this respect, some authors believe that the disparity between teachers' and students' conceptions of assessment may be caused by inconsistencies between teachers' conceptions and assessment practices, with students' conceptions primarily relating to their teachers' assessment practices (Borko et al., 1997; Remesal, 2006). The exceptions to this general trend are the studies by Brown (2008) and Brown et al. (2009b) showing consistency between teachers' and students' conceptions.

Comparing Teachers' Assessment Conceptions and Assessment Practices

Researchers have shown that the importance of studying beliefs and conceptions is their predictive relationship with practices (Barnes et al., 2015). In the domain of assessment, authors like Brown (2008), Brown et al. (2009), and Vandeyar and Killen (2007) are of the view that teachers' conceptions influence their decisions and professional activities. These authors believe that different assessment conceptions lead to different assessment practices. For example, teachers who conceive of assessment as important for improving teaching and learning will use formative methods of assessment, while teachers who have a conception of assessment for accountability will use summative assessment methods (Vandeyar and Killen, 2007).

Dixson and Worrell (2016) and Siarova et al. (2017) provided a set of characteristics of formative and summative assessment in classroom settings. AoL, also known as summative assessment, has the purpose to evaluate learning outcomes, provides information about student performance, serves to select or group students, and certifies learning and award qualifications. The methods used are projects, performance assessments, portfolios, papers, in-class examinations, standardized tests and national tests. Usually this is done by teachers and students are not active participants in assessment processes. These assessments include mostly closed questions, but they also use extended response items to evaluate how students apply their conceptual understanding and how they think critically, with the final goal of knowing how much a student knows. Summative assessments are graded, not frequent, and occur at the end of segments of instruction.

In contrast, Dixson and Worrell (2016) and Siarova et al. (2017) consider that AfL, or formative assessment, aims to improve students' learning, providing information to teachers and students to be used as feedback to modify teaching and learning. Thus, formative assessment is not usually graded. It can occur in two different practices: spontaneous—for example,

question-and-answer during instruction in real time—or planned, and it includes activities such as quizzes and homework exercises to assess student progress. Teachers have a key role in providing feedback and information about students' performances, yet the learner is also an important actor in the assessment process. Assessment tools used by teachers, such as observations, homework, feedback sessions, peer tutoring, self-assessment, question-and-answer sessions, comprehensive approaches to teaching and learning, student self- and peer-assessment, and effective feedback are frequent. Formative assessment occurs inside the teaching and learning process. The tools support deep learning, develop critical thinking, and promote students' interaction and continuity of the learning experiences (Dixson and Worrell, 2016).

Likewise, in a study with elementary and secondary teachers from Hong Kong, Brown et al. (2009) showed that practices of assessment to improve teaching and diagnose students' learning needs were predicted by the conception that assessment is about improvement. Practices related to preparing students for examination were predicted by the conception that assessments make students accountable. In contrast, findings from other studies (e.g., James and Pedder, 2006; Azis, 2015) have suggested that beliefs and practices of assessment are not related. Azis (2015), who studied the assessment practices and perceptions junior high teachers, noticed a conflict between practices and conceptions caused by the policy requirements of the existing assessment system in Indonesia. Teachers believed that the purpose of the assessment was to improve teaching and learning and to demonstrate the accountability of the students and the school. However, they felt that the state-wide examination policy requirements constrained their efforts to use assessment for these purposes. Hence, teachers' expectations of assessment and government policy were not aligned, causing a conflict between teachers' beliefs and assessment practices.

Similarly, James and Pedder (2006) found among English teachers that participants placed high value on AfL but their practices reflected a greater performance orientation. The authors posited that these results are caused by the testing context in England that required teachers to engage in performance-oriented practices and drive students to achieve in tests. Hence, the value attributed to summative assessment (traditional tests) in teachers' practices is much higher than the value ascribed to this modality in their conceptions when alternative modalities of assessment are highlighted.

These results show that teachers' conceptions are not always consistent with their assessment practices. The relationship between beliefs and practices is real but very complex, and these two elements influence one another (Opre, 2015), depending on individual and contextual factors that interrelate in accordance with each assessment situation (Barnes et al., 2015; Buehl and Beck, 2015). This congruence or incongruence between conceptions and practices has to be taken into account as it has different consequences for teachers' behaviors. According to Buehl and Beck, (2015) teachers' pleasure and wellbeing can be affected by a misalignment and, in extreme situations, teachers may even abandon their profession or implement inappropriate pedagogical practices.

Purposes of the Study

Authors like Suurtamm et al. (2010) suggest that we need studies to analyze how new ideas about assessment (e.g., AfL) are conceived of by teachers and students and how they are implemented in classroom practice. There have been policy changes within the Portuguese assessment guidelines in the past few decades, some reinforcing an assessment mode that we could call AoL, and sometimes supporting AfL. In the case of mathematics, Nortvedt et al. (2016) observe that, in actuality, assessment guidelines in Portugal are in line with those indicated in international terms (Mullis and Martin, 2015). That is, the regulations emphasize AfL, with a regulatory function over the teaching and learning process, and focus on assessing what is relevant in mathematics—not only what is easy to assess, but also the diversity of forms of assessment (Santos, 2004). Portugal is what the literature calls a low-stakes accountability context for assessment (Barnes et al., 2017). Nevertheless, three years ago there was a proliferation of national exams throughout schooling and frequent summative assessments to motivate students and inform parents, teachers, and schools. Nortvedt et al. (2016) found that, in the Portuguese context, there is a big gap between the curricular guidelines in mathematics and teachers' practices.

Brown (2011) states that teachers develop or adopt conceptions of assessment aligned with their own policy or legal frameworks. So, if teachers' beliefs are related to policies in their professional environment (Brown et al., 2011), teachers in Portugal should mainly possess a conception of assessment that guides improvements in teaching and learning, but also a conception that such assessment serves to judge the quality of student learning (student accountability). Hence, it was important to explore the conceptions that elementary teachers and students currently have about assessment and investigate whether those conceptions are aligned and similar to teachers' assessment practices.

This article presents data on Portuguese students' and teachers' conceptions of assessment and their practices in the domain of mathematics. Based on the research that revealed that teachers' conceptions of assessment differ across contexts and "reflect teachers' internalization of their society's cultural priorities and practices" (Barnes et al., 2015, p. 284), the present study intends to explore Portuguese primary school teachers' and students' conceptions of assessment and teachers' assessment practices and how they are related. Therefore, our research questions were:

1. How do primary teachers and students conceive assessment? Do their conceptions differ?
2. How do these primary teachers assess their students' learning?
3. What are the similarities and disparities that emerge when primary students' and teachers' conceptions and teachers' assessment practices are compared?

The Context of the Study: Assessment in Portugal (First Cycle)

In Portugal, basic education is compulsory and free. Children have to attend a public or a private school from the age of six

years. This level of education is divided into three cycles. We focused on the first cycle because there are few studies focusing on assessment conceptions in elementary school. This first cycle has a duration of four years and the components of the curriculum are articulated in a global manner through Grades 1 to 4. The process of learning and teaching is the responsibility of a single teacher. Assessment is predominantly informal and formative, and assumes a continuous and systematic character aimed at assisting teachers in obtaining all information necessary to implement pedagogical differentiation. Summative tests of educational progress (*provas de aferição*, external summative assessment) take place at the end of the second grade in mathematics and in the Portuguese language. Their purpose is to monitor the development of the curriculum in different areas and promote timely pedagogical interventions directed to the specific difficulties of each student. Yet seems to us to be an accountability purpose to the system and to the school, and results are reported to parents, teachers, and schools. Internal summative assessment occurs at the end of each trimester, with the purpose of classifying and certifying student progress or retention (Decreto-Lei 55/2018 de 6 de julho, 2018).

METHODOLOGY

Participants

Four schools were selected for this study, based on purposive sampling (Etikan et al., 2016). The main reason for choosing these schools was the possibility of experimental mortality. Data collected for this multiple-case study were part of a broader longitudinal research project that intended to understand the effects of teachers' assessment on students' achievement, motivation, and emotions. For the purpose of this project, it was necessary to ensure that students remained in the same school with the same teacher two years. This prerequisite was taken into account when selecting schools for this study. Once the schools were selected, our data were collected over two years. The data presented in this study refer to the first year.

Five teachers teaching third grade classes (A, B, C, D and E) and 82 students (between 11 and 23 students per classroom) participated in this study. Teachers (one male and four females) had between three and 25 years of experience. Students were aged 7–10 years (M = 8.07, SD = 0.34); 47 were boys and 35 girls.

Data Collection and Analysis

The data were gathered through individual interviews with teachers and focus group discussions with students, both of which were held at the end of the school year. Classroom observations and documents produced by the students (worksheets and tests) were also analyzed to determine teachers' assessment practices. Data collected were related to the domain of mathematics, a core subject in school education, which has high failure levels among Portuguese students (Organization for Economic Co-operation and Development, OECD, 2016).

Procedures to Classify Conceptions of Assessment

In order to understand teachers' and students' conceptions of assessment, one of the authors conducted semi-structured individual interviews with the teachers while another conducted focus group discussions with groups of four to five students. A total of 16 focus groups were conducted (two groups for Class D, three groups each for Classes B and A, and four groups each for Classes C and E). Groups were mixed gender and were defined through random sampling. The focus group started with the researcher trying to create a friendly environment, explaining the purpose of the study and giving time for the students to ask questions. The moderator ensured the participation of all members and kept the discussion informative rather than argumentative. The objective was not to reach a consensus but to collect all students' opinions.

Both interview and focus group questions were based on the literature (Remesal, 2006; Azis, 2015) and addressed five assessment topics for the teachers and four for the students (see **Table 1**). The same interview protocol was used in all the interviews and focus groups to ensure methodological consistency and control for reliability (Cohen et al., 2008). The content validity and appropriateness of the interview questions were verified by an expert in educational psychology.

All individual interviews and focus group discussions were audio-recorded and transcribed verbatim. Two of the authors performed a content analysis using the software MAXQDA 18. The interview and focus group contents were coded into fragments describing different categories, which were defined using both deductive and inductive approaches. Starting from the categories previously described by Brown (2008) and Harris and Brown (2009), the categories were progressively redefined through a cyclical process in order to fit and be representative of the reality of our data (Miles et al., 2014). Four qualitatively different categories of assessment conception were identified: external reporting, students' accountability, external motivation of students, and improvement of learning and teaching (see **Table 2**).

Intracoder consistency was assessed six months after the first analysis, with 84.6% agreement (mean κ for the categories was 0.925, between 0.894 and 0.953). For intercoder consistency, a second coder, working as a supervisor, confirmed the analyses of the first coder. Deviations from the initial analysis were discussed with all authors until final agreement or eventual recategorization.

Procedures to Classify Assessment Practices

Data about the teachers' assessment practices were gathered through the video recording of all the lessons in two learning units in the mathematics domain (one in the winter about stem and leaf diagrams and one in the spring about addition and subtraction with decimal numbers). We videotaped the teachers as they delivered regular lessons in the selected units, which varied in number (between three and eight sessions) and in duration (between 30 and 120 min). We also gathered all documents produced by the students during the lessons (worksheets, textbooks, notebooks, etc.). A total of 24 lessons

TABLE 1 | Topics and questions of the semi-structured interview protocol.

Topics	Questions for teachers	Questions for students
A. General assessment definition	For you, what is assessment?	You get assessed in school. If you had to explain to someone what assessment is, what would you say?
B. What to assess	What do you assess in math in grade 3 in these different situations? Is the assessment same or different? If it is same, what do you assess? If it is different, what do you assess in each type of situation?	
C. The intent of assessment	Why do you assess your 3rd graders? What is the need for these math assessments?	Why is your teacher assessing you in math? What is the assessment for?
D. Assessment process	Who does the assessment in mathematics? When do you assess? What assessment methods do you use with your 3rd graders?	Who does the assessment in mathematics? When does your teacher assess? How does your teacher assess you?
E. Use of the information from assessment	How can you and your students use the information from the different assessment tasks in maths? At the end of the trimester, what factors do you consider when assessing your students?	In the worksheets/tests does the teacher write anything? What do you do with the information you receive? In class when answering questions asked by the teacher, does she make comments? What kind of comments? What do you do with the comments the teacher makes to your answers during class?

TABLE 2 | Categories mentioned in the interviews and focus groups.

Categories	Examples
1. External reporting—assessment as a useful tool for reporting students' performance to parents, ministries, and schools.	"Because parents demand ... Parents demand." (T-E) "Only in the test ... classwork is to train us, and the test is for the teachers and parents to see if we know what to do." (S-A)
2. Students' accountability—assessment is used by students and teachers to evaluate students' performances, to indicate where they are in terms of learning and knowledge.	". . .When I conduct my assessment, it is, indeed, to measure their knowledge." (T-D) ". . .when you go ... you complete a worksheet, you show it to your teacher, and then, your teacher puts a right or a wrong mark." (S-A)
3. Extrinsically motivating students—assessment is described as a way of motivating students through competition, social pressure, or praise.	"...Throughout our lives, we are always being assessed, are we not? ... It is necessary for children to realize that, indeed, at some point, they have to be assessed..." (T-A) ". . .for the future, for when we go to work, ... imagine, if the teacher does not assess us, then we will not get a good job in the future." (S-A)
4. Improvement of learning and teaching—assessment is considered an important element in learning, knowledge, and teaching.	"What I assess on a daily basis is the evolution they are having: I Need to work with that boy more; he needs to do more training exercises." (T-C) "To see is we know the content, and if we have questions about it, she explains it to us." (S-E)

Note. *A, B, C, D, or E = Classes A, B, C, D, or E; S = Students; T = Teacher.*

were videotaped (1,767 min), and 860 documents (with 3,044 questions/exercises) were analyzed.

The qualitative data were analyzed using an observation grid derived from the literature about summative and formative assessment practices (Hattie and Timperley, 2007; Dixson and Worrell, 2016; Siarova et al., 2017; Singh et al., 2017). The dimensions observed are described in **Table 3**.

For the analyses of the dimensions oral questioning, type of question, who initiated the oral interactions, and oral feedback, we selected segments of video data for every teacher for closer analysis. These segments were the first moment of instruction, the last moment of individual work with textbooks or worksheets, and a moment when students worked in small groups. Two types of interaction (teacher initiation/student(s) response/teacher's feedback and student(s) initiation/teacher's feedback) were assessed for a maximum 30-min segment for each moment. In total, we examined 24 segments, being 680 min of footage, and observed 1,675 interactions. The proportion of these interactions did not differ significantly among the classes (z-test $p > 0.050$). Descriptive statistics of all the categories of the observed dimensions of assessment practices, such as relative percentages, were calculated to check through the data rapidly and protect against bias (Miles et al., 2014). These descriptive statistics were used to differentiate and characterize formative and summative practices for each of the dimensions observed, according to the theoretical basis described in **Table 3**.

Intercoder and intracoder reliability (four months apart) were excellent, with an observed agreement of 94.7% ($\kappa = 0.78$) and 94.5% ($\kappa = 0.80$) for intracoder reliability, and 91.5% ($\kappa = 0.72$) and 91.8% ($\kappa = 0.70$) for intercoder reliability.

RESULTS

The dataset obtained in this study was very large and complex in nature. Here we present a selection of the main analysis

TABLE 3 | Dimensions and categories of analysis for the assessment practices.

Dimensions	Categories
1. Who evaluates	1.1. Teacher assessment
	1.2. Peer assessment
	1.3. Self-assessment
2. Who initiated the oral interactions	2.1. The teacher, by asking questions
	2.2. The students, by asking questions or making statements to obtain feedback
3. Oral questioning: Who is questioned	3.1. All students (or almost all students) are questioned
	3.2. A few students (mostly the same ones) are questioned
4. Assessment tools used	4.1. Worksheets developed by the teacher
	4.2. Worksheets from a textbook
	4.3. Classroom observation
	4.4. Group work
	4.5. Homework
	4.6. Oral presentations
	4.7. Checking daily notebooks
5. Oral questioning: Type of question	5.1. Open
	5.2. Closed
6. Written questioning: Level of cognitive complexity	6.1. Low—remember (focus on retrieving relevant knowledge from long-term memory) and understand (focus on clarifying, recalling, naming, and listing)
	6.2. High—apply (focus on prior knowledge to solve a problem), analyze (focus on carrying out a procedure in a given situation) and evaluate (focus on making judgments based on criteria and standards)
7 and 8. Oral and written feedback (Focus)	7.1. Feedback at the self-level (praise or criticism without task-related information)
	7.2. Feedback at the task and product level (corrective feedback, pointing out errors or providing correct forms)
	7.3. Feedback at the process level, aimed at the processes used to complete the tasks (clarifications, hints, suggestions for the future, or asking for explanations)
	7.4. Feedback at the self-regulation or conditional level, which engages students' skills in self-evaluation (encouraging self-assessment)

TABLE 4 | Overview of students and teachers' assessment conceptions.

		Classes				
Criterion		**A**	**B**	**C**	**D**	**E**
Categories most often covered	Teacher	ILT	ILT	ILT	ILT	ITL
	Students	ITL	SA	SA	SA	SA
Categories where several different aspects were mentioned	Teacher	ILT	ILT	ILT	ILT	ER & SA & ITL
	Students	SA & ILT	SA	SA	SA	SA & EMS
Categories mentioned in (nearly) all the interview topics	Teacher	SA & EMS & ITL	ILT	ILT	SA & ILT	SA
	Students	SA & ILT	SA & ILT	SA & ILT	SA	SA

Note. *EMS = Extrinsically motivating students; ER = External reporting; ILT = Improvement of learning and teaching; SA = Students' accountability;*

categories. We first summarize teachers' and students' conceptions of assessment, and teachers' assessment practices. Then we compare and contrast their conceptions and teachers' assessment practices in more detail.

Teachers' and Students' Conceptions of Assessment

Teachers and students mentioned all four categories during the interviews. In **Table 4**, we summarize for each teacher and for their students the categories most often covered, the categories with richer content (i.e., several different aspects of the category were mentioned), and the categories mentioned in nearly all the interview topics. This information allowed us to order our participants along the continuum ranging from the AoL pole (with greater focus on certification and accountability) to the AfL pole (with greater focus on improving learning and teaching). Most of the students seemed to be predominantly at the AoL pole of the continuum, while most of the teachers were at the AfL pole (see **Figure 1**).

Teachers B and C mentioned the improvement of learning and teaching many times; they mentioned several different aspects of this category, and also mention it, throughout the entire interview (See **Table 4**). Teachers' B and C discourse indicated that their conception of assessment was largely focused on improving students' learning, closer to the AfL pole (see **Figure 1**). Teachers A and D had mixed conceptions, having highlighted assessment for learning as the purpose of assessment but also constantly mentioning throughout the interview the students' accountability and extrinsically motivating students as important purposes of assessment. Finally, Teacher E also had a mixed conception, yet it was closer to the AoL pole. Teacher E

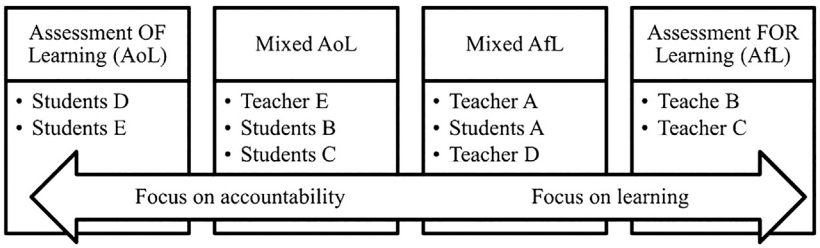


FIGURE 1 | Description of the Conceptions of Teachers and Students.

mentioned the improvement of learning and teaching more times than any category, but her discourse was very rich about external reporting and students' accountability, mentioning different aspects of these categories. Improvement of learning and teaching was mostly mentioned when taking about feedback, while students' accountability was mentioned throughout all topics of the interview (see **Table 4**). Therefore, this teacher also focused on improving teaching and learning, but emphasized external reporting and students' accountability as the primary purpose of assessment, while considering that only feedback had the purpose of improve learning and teaching.

In contrast, most of the students in all classes considered the primary function of assessment to be verifying the correct answer. Most of the comments related to the improvement of learning and teaching purpose were offered only when the topic of oral feedback was questioned, while students' accountability was mentioned throughout all topics (see **Table 4**). Students in classes B and C presented a mixed conception, but with a stronger presence of an AoL conception. Only the students in Class A presented a mixed conception more focused on AfL, mentioning the improvement of learning category several times; the category was broken down into several different aspects and mentioned in most of the interview topics. Students from classes D and E mentioned students' accountability more times, using a very rich discourse (they mentioned several aspect of this category) and mentioning it throughout the interview. Hence, we considered that students of Teachers D and E presented an AoL conception.

Teachers' Assessment Practices

The structure of the mathematics lessons delivered by these five teachers was uniform and followed the following sequence:

- Instruction: the teacher introduced content that students had not previously worked on by using expositions, demonstrations, illustrations, problem-solving and class discussion. In these moments, the most used assessment tool was oral questioning, used mostly for certification of previous knowledge.
- Practising the new content: students practised or applied the newly introduced content through individual or group tasks, mostly using worksheets developed by the teacher or from textbooks. In these moments, the assessment tool more used was classroom observation without keeping a record. The feedback used in these moments was more focused on the process, but most feedback was still at the level of the task.
- Formal assessment: Teacher assessed mostly by checking the answers to the exercises completed during the practice moments through oral or written feedback. Most of the feedback given in these moments was corrective.

Based on the analysis of the different moments of the lessons observed in the sample, all teachers presented mixed assessment practices (see **Table 5**), with a tendency to use summative assessment in more dimensions, closer to an AoL pole. Teacher E consistently presented very few AfL practices.

Analyzing Conceptions and Practices

This section presents a more detailed description of the data concerning teachers' conceptions of assessment and their practices. The objective was to find similarities and discrepancies between teachers' assessment conceptions and their practices and to reflect on the effects of teachers' conceptions and practices on students' conceptions.

Looking at Teachers B and C and Their Students

Though Teachers B and C both presented an AfL conception, they still showed some particular differences. One was related to the focus of the improvement. Teacher B was more focused on improving teaching and considered assessment important for modifying teaching strategies for the benefit of students (e.g., "If I see that a large majority of the group failed in a particular subject, then I see that this issue has to be... taken up in a different way because it was not assimilated as I thought it should be"). Teacher C focused more on improving students' learning through their self-regulation. For her, the purpose of assessment was to identify students' weaknesses, foster students' self-regulation skills, and provide for individual needs. This purpose was borne by this teacher's understanding of AfL: "The assessment process is discussed with them [students]... It is work that is done together, me and the students... they plan what they want to work on... after realizing the difficulties of the students, I try to meet the needs of each one."

These teachers also differed in their conceptions of feedback. When questioned about feedback, the teacher of Class C emphasized the importance of ensuring that the student understood and agreed with the assessment (e.g., "When there is something negative, I ask, 'Do you agree with what I wrote or

TABLE 5 | Overview of teachers' assessment practices.

	Description of teachers' assessment practices				
	Classes				
Dimension	**A**	**B**	**C**	**D**	**E**
1. Who evaluates?	100% Teacher	100% Teacher	60% Teacher +40% Peer assessment	100% Teacher	100% Teacher
2. Who initiated the interactions	39.4% Students	56.1% Students	45.6% Students	37.9% Students	29.8% Students
3. Oral questioning - who is questioned	33.3% All students	0% All students	50% All students	50% All students	20% All students
4.Assessment tools used	- Worksheets developed by the teacher - Group work - Observation	- Worksheets developed by the teacher and from textbooks, - Observation - Group work	- Worksheets developed by the teacher, - Observation - Oral presentation - Group work	- Worksheets developed by the teacher and from textbooks - Observation - Oral presentation - Group work	- Worksheets developed by the teacher and from textbook - Observation
5. Oral questioning - kind of question used	82.7% Closed	80.1% Closed	74.0% Closed	66.3% Closed	94.1% Closed
6. Written questioning Cognitive level	63.2% High	55.5% High	31.9% High	59.3% High	63.9% High
7. Oral feedback (Focus level)	5.5% Self 55.9% Task 38.6% Process	16.1% Self 52.1% Task 31.7% Process	6.3% Self 69.9% Task 23.8% Process	8.5% Self 50.8% Task 40.3% Process 0.4% Self-regulation	5.2% Self 71.4% Task 23.3% Process
8. Written feedback (Focus level)	0% Self 98% Task 2% Process	5.5% Self 80.0% Task 14.5% Process	2.7% Self 91.1 Task 6.1% Process	14% Self 84.1% Task 1.8% Process	1.8% Self 91.6% Task 6.6% Process

what I said?' and they actually say 'Ah! Yes, I could have tried harder. You're right'. And that's it. And indeed, then there is. . . a return. They make the effort to be more aware, more focused"). This teacher considered this agreement very important because students used that feedback to plan their learning (e.g., "In the following week, in their Individual Work Plan, they paid attention to everything that had been transmitted by me, whether oral or written"). Teacher B highlighted the use of feedback only for general encouragement (e.g., "For example, when there is a kid who is systematically failing an account, day by day, when he finally succeeds, I say 'Hallelujah!'")

Contrary to their assessment conception being very focused on the improvement of learning and teaching, the assessment practices of Teachers B and C were mostly mixed. Teacher C presented mixed practices, close neither to the AfL pole nor the AoL pole. Overall, it was she who evaluated the students, but she sporadically facilitated peer evaluation. She used alternative student-centred assessments (specifically, peer assessment) in each unit that was observed. She encouraged her students to initiate interaction related to the topics that they were studying (**Table 5**). She also used a wide variety of tools to acquire information about students' knowledge. All these practices are used when the teacher's purpose is to increase students' learning (Cizek, 2010). However, she posed several closed-ended questions at a low cognitive level and provided students with correct answers, using very summative feedback. The sheets and tests used by the teacher posed low cognitive level questioning (only 31.9% of the questions were from the applying or analyzing level), which do not help students increase autonomy (Singh et al., 2017). She largely provided task-related oral and written feedback, and this practice was incongruous with her beliefs that focused on students' self-regulation. There was little room for process-related feedback when the questions focused on low cognitive levels.

Teacher B's classroom assessment practices also diverged from her AfL conception of assessments. On the one hand, observations of her assessment practices confirmed that she was the only evaluator across all the lessons. Teacher B used closed-ended questions more frequently than open-ended questions, questioning only a few students—usually those who volunteered. Most of the oral feedback provided during this process focused on the task and on providing the correct answer (See **Table 5**). Furthermore, a significant percentage of the feedback also focused on the self (19.2%, the highest rate of all teachers), offering general encouragement to students, as indicated in the interview, but with little effect on students' learning (Hattie, 2012). Written feedback was mostly provided in relation to formal assessments for verification purposes, even though Teacher B provided a higher percentage of process-related feedback (12.9%) than other teachers. On the other hand, Teacher B created a supportive environment within which the students felt comfortable enough to express their thoughts and ideas. Indeed, this was one of the classes in which students tended to initiate more interactions (**Table 5**). Written questions primarily pertained to understanding and application cognitive levels, and some questions pertained to the analysis level. This was indicative of the use of high-level questioning. She used different classroom tools to collect information during individual and group work (questioning, observation, evaluation sheets, projects, and tests).

The influence of accountability on students' thinking was common among students from classes B and C; both showed a mixed conception of assessment, closer to the AoL pole. Students' conceptions from both classes were more consistent with their teacher's practices than with their teachers' conceptions. For these students, assessment primarily fulfilled a social function of certification of their knowledge and served to regulate learning minimally; as such, it appeared to be a process controlled mostly by their teachers. Students often stated that assessment was the process of checking their work (e.g., "Assessment is something that, when you do something wrong, your teacher says to you, 'Oh, this is not ok!', and when you understand and you do well the teacher says: 'Ah! Very well, now you understand'"). When questioned about the type of feedback provided by the teacher, they mentioned verification (at task level or the self-level), such as "good girl" or "great effort." This was coherent with the type of feedback they received from their teacher. The triangulation of our findings from teacher interviews, classroom observations and student focus groups revealed that these teachers' vision of assessment was for improvement, even though their actual assessment practices (especially feedback practices) were coherent with students' conceptions that assessment should hold students accountable for learning.

Looking at Teachers A and D and Their Students

Teachers A and D presented a mixed conception, closer to the AfL pole. While assessment fulfilled a primarily formative function for these teachers, it was also considered important to establish common minimum levels that all students must achieve. In their interviews, these teachers focused mainly on the improvement of learning and teaching [e.g., "I assess the students every day, their evolution, you see? Then I know, 'I need to work more with that student' (or) 'He needs one more exercise, more training'" – Teacher D]. Based on students' performance in evaluation tests, the teachers realized that some content had not been learnt well and tried to fill this gap by adjusting their teaching and providing learning guidance. Nevertheless, a certification purpose was present in their conception of assessment (e.g., ". . . until some day, they had to know, and that's it. They really have to know. If they do not know, they have a 'wrong'" – Teacher A). Teacher A highlighted the importance of assessment for students' academic and professional futures, for preparing students for future assessments, and for developing the skills needed in "real life":

> "... thus, throughout our life, we are always being assessed, aren't we? By other people's opinions or by our performance at work. We are always being evaluated. Children need to realize that, sometimes, they have to be assessed in a more formal or informal way. It is important that they learn how to react when they are assessed, isn't it? (Teacher A)."

In this regard, assessment is considered a pre-requisite for students to be prepared for the social challenges of everyday life. In turn, Teacher D emphasized the verification of students' current knowledge (e.g., "When I use assessment it is to appraise their knowledge") and feedback consisted of information to students about progress, but at the task level (Hattie and Timperley, 2007) (e.g., "I think feedback should be used to tell my students how well they are performing in school and if what they are doing is correct or not").

Teacher A presented mixed assessment practices, closer to the AoL pole. Here, the teacher controlled all the assessment processes. She was the only evaluator across all the lessons that were observed. Most of the interactions were initiated by the teacher. The main assessment mode used by this teacher was oral questioning, addressed to the entire class, but only a few students participated, and it was the same students who typically volunteered to answer. Most questions were closed-ended, which seemed to indicate that the purpose of these questions was to evaluate students' level of understanding of the unit content. In contrast, the written questioning used by the teacher concentrated on the use of high-level questioning, but Teacher A neither checked the work completed by the students nor worked out the solutions on the blackboard, which meant that little feedback was offered to the students. The little feedback offered was almost always focused on the task, although some oral feedback also focused process (38.6%; see **Table 5**). In sum, Teacher A evaluated only certain aspects of the learning process, evaluated students' prior knowledge, and corrected errors. These summative practices hold students accountable for the concepts that they learn.

However, students' and teachers' in Class A were aligned in their mixed conception closer to the AfL pole, which was inconsistent with the teacher's greater accountability practices. These students highlight the importance of assessment for their learning and teaching (e.g., "The assessment is for the teacher to help us when we have some difficulty. For example, the teacher provided some extra classes, talked to our parents to explain the situation, assigned more homework, or spent more time to try to find a solution... That's what the assessment is for: to know what we know, what we do not know, and what our difficulties are"). Just like their teacher, students in this class often stated (more than those of other classes) the importance of assessment for their future (e.g., assessment is ". . .for the future, for when we go to work... imagine, if the teacher does not evaluate us, in the future we do not have a good job"), reflecting ideas expressed by the teacher during the interview.

Teacher D valued the formative and summative purposes of assessment and sought to achieve synergy during the assessment process. This conception of assessment was consistent with her assessment practices. In Class D, assessment was controlled by the teacher, and most of the interactions established between teacher and students, two-thirds of the observed interactions, were initiated by the teacher. Furthermore, the oral questions that were posed to the students were largely closed-ended. However, one-third of them were open-ended questions and the teacher used them to promote student learning. The feedback provided to the students (written and oral) focused largely on the task. However, oral feedback at the process level was provided to a slightly greater extent to support and guide students. Moreover, Teacher D was the only teacher that offered feedback at the self-regulated level (See **Table 5**). To gauge and guide student learning, several different assessment tools were used.

In contrast to these more formative practices and to their teacher's conception, students presented a conception very close to the AoL pole. Only one focus group mentioned learning as a purpose of assessment, but with the final purpose of "progressing well to the fourth grade." The high degree to which students in Class D endorsed the accountability purposes of assessment seems to be more consistent with some summative assessment practices embedded in their teacher's activities. Teacher D checked the work completed by each student at the end of almost every lesson and provided written feedback (which focused largely on the task and aimed to check the correctness of students' answers). She always provided the final written evaluation (a symbol, which indicated that the answer was correct) only after the students successfully completed their task: hence all the tasks completed by the students were marked as having been completed correctly. Consequently, for these students, assessment happened when the teacher checked the answer, which is mostly necessary to set minimum standards that all students must meet to be promoted to the next school year (e.g., "I think it serves so we feel… so we know, that we did 'good'… that we did a good job and… that we are ready for fourth grade and so on."). These students had little concern for improving learning. It was more important that all students achieved the pre-defined objective: "Assessment is important to the teacher to know if we have mastered the contents so that we can progress to a higher grade."

Looking at Teacher E and Her Students

Teacher E presented a mixed conception of assessment. When talking about assessment, Teacher E was focused on the certification of learning, specifically external reporting [e.g., assessment is done because "parents required (me) to do it"] and students' accountability (i.e., the main purpose is to summarize student achievement: "…for me, assessment is done based on percentages attributed to each performance criterion: 30% from formal assessment and 70% from informal assessment"). On the other hand, for this teacher, the feedback should be focused on the process for it to be an important tool for teaching and learning ("The main goal in assessment is that children realize what they actually did wrong… that students engage in error correction strategies following error detection and that they strive to improve their learning. The important thing is for them to understand that if they fail, they can seek help").

Teacher E presented mostly AoL assessment practices. Teacher E played a substantial role in student assessment. She was the only evaluator across all the lessons, and classroom interactions were dominated by this teacher: only one quarter of the interactions were initiated by the students. Almost all the oral questions were closed-ended, and her feedback focused largely on the task. She did provide process-related feedback, but only when oral questioning was conducted (See **Table 5**). Written questions pertained to the lower cognitive levels. There was no concern on the part of the teacher about incentivizing all students to participate actively in classes. The tools that this teacher used to evaluate students' learning (questioning, observation, textbook exercises, worksheets and tests) were designed to be used individually. Teacher E checked the work completed by each student at the end of almost every lesson and provided written feedback to verify the correctness of their answers.

Teacher E's practices and conceptions were reflected in the students' conceptions, presenting an AoL conception. Students emphasized that assessment was mostly necessary for assigning grades, categorizing students and determining if students can be promoted to the next grade. Both teacher and students mentioned the importance of assessment to motivate students to achieve the "honor roll," a prize given for the best students (e.g., "The assessment allows us to get awards and go to the honor roll"). Students also consider that without assessment, they will not even try to learn (e.g., students mention that "if there is no assessment, there is no point for me doing the worksheet," while the teacher said: "They know that this trimester they will not be assessed in the science class, so the kids are totally careless… they are not that careful as when they know they will be assessed—'I will be evaluated, I need to pay some attention'").

DISCUSSION

In the present study, we first aimed to investigate the classroom conceptions of five primary teachers and their students' assessment conceptions. Our analysis resulted in four categories which ranged from completely focused on accountability to focused on learning. There was variation in how teachers and students conceived assessment. Nevertheless, our analysis revealed no consistency between the conceptions of students and teachers except for Teacher A and her students.

On the one hand, the findings revealed that teachers believed assessment was mainly intended to improve learning and teaching (pole AfL). These results are similar to those obtained by Remesal (2009, 2011), where primary teachers revealed a pedagogical conception of assessment, in extreme and mixed forms. These outcomes are also in line with the Portuguese assessment guidelines (Decreto-Lei 55/2018), which indicate that assessment in the first cycle of schooling should help teachers and students to improve teaching and learning—that is, help students to perceive what they should improve upon and how, and help teachers adjust their pedagogical strategies to students' needs. This conception meets Black and Wiliam (2018) definition of formative assessment. These findings are important because, according to the studies of Brown (Brown, 2008; Brown et al., 2009b), teachers' conceptions influence their decisions in the classroom. Given this argument, it is expected that these teachers will use formative assessment approaches and techniques to better understand students' learning needs and adjust their teaching strategies to promote students' achievement.

On the other hand, the results also illustrated that the students' conceptions of assessment stood at the AoL pole (extreme or mixed forms), with a strong emphasis on summative assessment. Only the students in one class (Class A) revealed a mixed conception closer to AfL, similar to their teacher's conception. These results provide evidence that, from the time these children went to primary school, they agreed less with an improvement conception of assessment, in contrast to the results obtained by

Harris and Brown (2009). These results are important because some researchers (Brown and Hirschfeld, 2007; Harris and Brown, 2009) state that students' conceptions of assessment guide and determine how they study. Therefore, our findings reflect that for these young students (about eight years old) assessment is "high stakes" in driving their study behavior toward grades. There is a clear and unambiguous consequence to students based on their grades. These beliefs are congruent with the purposes of summative assessment, in that assessment should be used to measure students' learning at the end of a unit, to promote better learning outcomes, to get a certification for school completion, or to select students for entry into further education (Organization for Economic Co-operation and Development, OCED, 2008).

When comparing the conceptions of these teachers and students, our results differed from those of Brown (2008, 2012) and Brown, Irving, et al. (2009), and were similar to the results from Remesal (2006). Probably, as observed by Borko et al. (1997), the disparity between teachers' and students' conceptions could have resulted from the discrepancy between teachers' conceptions and their practices. Indeed, our results showed an inconsistency between teachers' conceptions and practices, and more coherence between teachers' practices and students' conceptions. This allows us to think that these teachers' assessment practices may in some way contribute to the way their pupils conceive the assessment process. This reinforces the socio-constructivist point of view that conceptions of assessment are social constructs that depend on the pedagogical experience and the environment in which teachers and students are involved (Gipps, 1999).

Our analysis also focused on teachers' assessment practices and on the relationship between teachers' conceptions of assessment and their assessment practices. Our analysis revealed low consistency between conceptions and practices, though there was one exception, Teacher E. Most teachers use summative assessment practices and emphasize the measurement of learning and control of the assessment process, using feedback at the task level. Furthermore, in the observed lessons and documents, the teachers provided students with little effective oral feedback and when it did occur it was mainly at the task level. Still, teachers rarely provided written feedback at process level; they mostly gave grades that did not reveal the real needs of students. Thus, the results of the present study showed that these teachers used assessment mostly to measure the reproduction of knowledge (most questions were from a low cognitive level). The teachers probably wanted to ensure that their students reached a level of success or proficiency necessary to enter the second cycle of studies.

Accordingly, based on the statement that assessment guides students' learning and competences for learning, our results suggest that the assessment practices of these teachers should be carefully considered. Peer assessment was mentioned by only one teacher and self-assessment was not mentioned at all. However, these assessment methods are increasingly important for dialogical teaching and learning, where formative assessment takes on a very relevant role. The assessment practices applied by the teachers participating in the present study seem not to guide students' learning.

The results of the present study indicate that, at least in circumstances such as those observed in these five teachers, the teachers' conceptions are not consistent with their teaching practices. In our study, only one teacher (E) showed consistency between conceptions and practices. She conceived assessment as a mean of measuring factual knowledge and she adopted mostly summative assessment practices.

Contrary to the results obtained by Brown (2008), Brown, Irving, et al. (2009), and Vandeyar and Killen (2007), which concluded that the conceptions that teachers have about assessment influence their practices, we found in the present study that their practices did not always reflect their beliefs. Some factors may explain these discrepancies between our results and past findings.

One set of explanations relates to the nature of the methods used. The current study specifically focused on qualitative data, while Brown's studies (2008, 2012; Brown et al., 2009a) worked with self-administered questionnaires with closed-ended rating scales and statistical data. There are strengths and weaknesses to these two crucial research paradigms in education, qualitative and quantitative. Nevertheless, we highlight the benefits of using qualitative research in the assessment domain. Qualitative approaches allow us to achieve a more profound understanding of the data gathered in all phases of the process (Rahman, 2017); it is easier to understand the behavior of the participants, the interviewees, and the contextual and socio-cultural influences on the behavior of participants during interviews. Of course, there are some limitations, such as small sample size, which make the results unreliable and ungeneralizable and hence not preferred by policymakers (Rahman, 2017).

Another explanation is related to the Portuguese guidelines on assessment. The actual assessment policy regulation in Portugal states that teachers should certify their students' knowledge at the end of each trimester for parents, students, and the school (Decreto-Lei 55/2018). This purpose was also mentioned a few times in teachers' conceptions. It is clear that assessment in Portugal involves a relationship between formative and summative purposes, with an evident emphasis on improvement for learning and teaching. So, it is understandable that teachers think assessment should inform teachers about the changes they have to implement in teaching, and should inform students about their strengths and weaknesses and help them reorganize their learning in the future. But why, then, do teachers not implement more formative assessment? We can presume that the teachers' conceptions reflect the obligations of the National Policy on Assessment in primary schools in Portugal (the emphasis is on formative assessment). However, the constraints imposed by the context on the materialization of teachers' assessment conceptions (Opre, 2015) can result in disparities between conceptions and practices.

An additional possible explanation for such discrepancy is the historical context of assessment in Portugal. The certification of students was the main assessment purpose during recent years (2011–2016): in this period, examinations proliferated at all levels of schooling, with various consequences for teachers and students. We believe that, despite new regulations, the current assessment practices of these teachers are still embedded in the

assessment purposes that dominated the assessment system until recently—the purposes of student and school accountability.

It is also possible to explain the discrepancies between conceptions and practices through the pressures of high-stakes assessments on teachers' work (Brown et al., 2009a). Although, since 2016, there has not been a national maths exam at the end of the fourth grade of schooling in Portugal, national tests are still held in the second, fifth, and eighth grades. These exams are intended to detect students' difficulties in these middle grades and help teachers find strategies that help them overcome these difficulties in the following years. However, results serve a range of purposes, including evaluative feedback to teachers, schools, and the national system about the effectiveness of students' performance. These results also serve to advise teachers and parents about decisions on future study strategies (Despacho normativo 1-F/2016 de 5 de abril, 2016). When we interviewed these teachers, their students were in the third grade, some months after having performed those exams. We can suppose that this type of assessment has an effect on these teachers' assessment practices and on these students' conceptions, which reinforces the accountability purpose of assessment.

Finally, another factor to take into account when evaluating teachers' conceptions is the fact that they may have given socially desirable responses that differ from what they actually do in the classroom. When faced with questions about assessment, which can be a sensitive topic for them, teachers may answer in accordance with what they believe is socially expected (Eivarsen and Våland, 2010). These concerns may remain, even though participants were repeatedly assured of confidentiality and several strategies were adopted to reduce social desirability bias (Ananthram, 2016): for example, teachers themselves volunteered to participate in the interview, they remained anonymous, and we provided a brief overview of the study goals.

What we can infer here is that the relationship between conceptions and practices is complex, and that individual, social, and contextual factors could influence one another with implications for teaching and learning. We can also assume that, in some circumstances, teachers' assessment practices are closer to students' conceptions of assessment (Opre, 2015) than to their own conceptions. The outcomes showed that these young students perceived assessment mostly in the form of a grade: this view of assessment can bring some obstacles to an assessment conception that mediates their learning and achievement, and can have consequences for how they involve themselves in assessment tasks. According to Black and Wiliam (2018), assessment supports learning when students receive feedback that takes learning forward. In our study, we observed that in several situations teachers provided evaluative, general, written, and oral feedback, frequently focusing on results that reinforced students' conceptions of AoL. So, assessment, in terms of students' conceptions and teachers' practices, is intended to serve both certification and improvement of teaching and learning, but priority is given to student and school accountability.

In four of the cases presented, teachers' practices and conceptions seem to be generally inconsistent. Cognitive dissonance theory (Festinger, 1957) suggests that individuals seek to maintain consonance among multiple cognitions of beliefs and behavior. The theory adds that if there is dissonance between beliefs and behavior, "individuals engage in changing their beliefs and/or behaviors to make them consonant in order to achieve cognitive consistency" (Guerra and Wubbena, 2017, p. 39). Therefore, to reduce the dissonance observed in the present research, we think that teachers need to focus primarily on enacting changes in their practices rather than changes in beliefs. When applied to the present study, we suppose that teachers' belief-practice inconsistency is likely to be related to the policy context of a high-stakes test-influenced environment. The intense pressure upon teachers comes from focusing only on high-stakes testing and, in some circumstances, can lead to burnout (Pishghadam et al., 2014). Although the existing literature reinforces that beliefs shape teacher behavior (Karaagac and Threlfall, 2004), in the present research, it seems that the teachers' goals drove their behavior more than their beliefs did.

If teachers are not motivated to change their classroom assessment practices, the conflict with their beliefs will remain evident, though the teachers were aware of that inconsistency, as it was stated by some of them. It may be expected that those teachers who considered assessment inaccurate, neglected, or unfair may become indifferent and unmotivated toward their learners and their profession (Pishghadam et al., 2014). Additionally, teacher burnout can reduce students' intrinsic motivation, which may reduce learning (Shen et al., 2015).

Our results tend to confirm Remesal's (2009, p.49) hypothetical model that "young pupils perceive assessment practices, whichever form they take." From an interpretivist approach, it is important to recognize the complexity of interactions among students, teachers, and assessment (Gipps, 1999). Students' views of assessment are conceptualized and reconstructed through their experiences within the social setting of the classroom. So, individual factors such as the learner's expectations of the classroom process, their interpretation of the demands of the task, and the criteria for success are in constant relationship with social factors such as teachers' expectations and their pedagogical practices. Hence, if assessment practices are associated mostly with learning certification, students may develop a more passive role in their learning process (Remesal, 2009).

Final Considerations and Educational Implications

Our most striking finding was that the assessment practices in the study context were mostly traditional (summative) and that most academics described the purpose of assessment in a dialogical way, emphasizing formative assessment and the importance of feedback for learning or to modify teaching strategies and adapt them to students' specific needs.

In order to promote significant learning in these students, we think it is necessary to introduce changes that make their teachers' assessment practices authentic and more formative, consonant with their conceptions. Our results indicate that assessment practices change slowly. The ways of thinking (in

line with the legislation on student assessment) and practising differ in teachers. Therefore, we suggest that if we want a more dialogical teaching and learning process, more specific research in real assessment contexts is needed to understand teachers' assessment practices. We suggest that the development of assessment practices could be supported through more collaborative practices of assessment. Sharing positive experiences of assessment in collaborative settings may result in higher awareness of the relationship between assessment conceptions and practices (Siarova et al., 2017).

Thus, if learning is socially situated, the role of teachers in analyzing and reflecting on the needs of their students requires that emphasis be placed on formative assessment of pupils' understanding. So, it is important that teachers and students come to a common understanding of the meaning of communicated feedback in order for students to understand how to improve their achievement (Gipps, 1999; Andersson, 2016). In this sense, newer practices of assessment, deriving from the socio-cultural approach, are required. The assessment task should highlight how the learning process is developing, and has to be understood as an interactive, dynamic, and collaborative task (with the teacher and with the peer group) in order to develop students as self-regulated learners (Gipps, 1999).

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 Ethics Committee of ISPA. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

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

FUNDING

This study was supported by the FCT - Science and Technology Foundation - PTDC/MHC-CED/1680/2014 and UIDP/04853/2020

ACKNOWLEDGMENTS

We wanted to say thank you to Marta Gomes and Cristina Sanches for their contributions to this project. It was helpful to have someone with whom to discuss ways of fine-tuning and optimizing our processes. Our sincere thanks for your invaluable assistance.

REFERENCES

Ämmälä, A., and Kyrö-Ämmälä, O. (2018). Conceptions of school assessment: what do Finnish elementary school students think of assessment? *Educ. North* 25 (1–2), 275–294. doi:10.26203/0g1s-hf54. Available at: https://www.abdn.ac.uk/education/research/eitn/journal/558/ (Accessed September 11, 2019).

Ananthram, S. (2016). "HRM as strategic business partner: the contributions of strategic agility, knowledge management and management development in multinational enterprises – empirical insights from India," in *Asia Pacific human resource management and organizational effectiveness*. Editors A. Nankervis, C. Rowley, and N. M. Salleh (London, United Kingdom: Elsevier), 87–109. doi:10.1016/B978-0-08-100643-6.00005-1

Andersson, N. (2016). Teacher's conceptions of quality in dance education expressed through grade conferences. *J. Pedagogy* 7 (2), 11–32. doi:10.1515/jped-2016-0014

Azis, A. (2012). Teachers' conceptions and use of assessment in student learning. *Indones. J. Appl. Ling.* 2 (1), 41–45. doi:10.17509/ijal.v2i1.72

Azis, A. (2015). Conceptions and practices of assessment: a case of teachers representing improvement conception. *TEFLIN J.* 26 (2), 129–154. doi:10.15639/teflinjournal.v26i2/129-154

Barnes, N., Fives, H., and Dacey, C. M. (2015). "Teachers' beliefs about assessment," in *International handbook of research on teachers' beliefs*. Editors H. Fives and M. G. Gill (New York, NY: Routledge), 284–300.

Barnes, N., Fives, H., and Dacey, C. M. (2017). U.S. teachers' conceptions of the purposes of assessment. *Teach. Teach. Educ.* 65, 107–116. doi:10.1016/j.tate.2017.02.017

Black, P., and Wiliam, D. (2018). Classroom assessment and pedagogy. *Assess. Educ. Principles, Policy Pract.* 5 (1), 7–74. doi:10.1080/0969594X.2018.1441807

Borko, H., Mayfield, V., Marion, S., Flexer, R., and Cumbo, K. (1997). Teachers' developing ideas and practices about mathematics performance assessment: successes, stumbling blocks, and implications for professional development. *Teach. Teach. Educ.* 13 (3), 259–278. doi:10.1016/s0742-051x(96)00024-8

Brown, G. T. L. (2004). Teachers' conceptions of assessment: implications for policy and professional development. *Assess. Educ. Principles, Policy Pract.* 11 (3), 301–318. doi:10.1080/0969594042000304609

Brown, G. T. L. (2008). *Conceptions of assessment: understanding what assessment means to teachers and students*. New York, NY: Nova Science Publishers.

Brown, G. T. L. (2011). Teachers' conceptions of assessment: comparing primary and secondary teachers in New Zealand. *Assess. Matters* 3, 45–70. doi:10.18296/am.0097

Brown, G. T. L., and Harris, L. (2012). Student conceptions of assessment by level of schooling: further evidence for ecological rationality in belief systems. *Aust. J. Educ. Dev. Psychol.* 12, 46–59.

Brown, G. T. L., and Hirschfeld, G. H. F. (2007). Students' conceptions of assessment and mathematics: self-regulation raises achievement. *Aust. J. Educ. Dev. Psychol.* 7, 63–74.

Brown, G. T. L., Irving, S. E., Peterson, E. R., and Hirschfeld, G. H. F. (2009a). Use of interactive-informal assessment practices: New Zealand secondary students' conceptions of assessment. *Learn. Instr.* 19 (2), 97–111. doi:10.1016/j.learninstruc.2008.02.003

Brown, G. T. L., Kennedy, K. J., Fok, P. K., Chan, J. K. S., and Yu, W. M. (2009b). Assessment for student improvement: understanding Hong Kong teachers' conceptions and practices of assessment. *Assess. Educ. Principles, Policy Pract.* 16 (3), 347–363. doi:10.1080/09695940903319737

Brown, G. T. L., Lake, R., and Matters, G. (2011). Queensland teachers' conceptions of assessment: the impact of policy priorities on teacher attitudes. *Teach. Teach. Educ.* 27 (1), 210–220. doi:10.1016/j.tate.2010.08.003

Brown, G. T. L., and Remesal, A. (2017). Teachers' conceptions of assessment: comparing two inventories with Ecuadorian teachers. *Stud. Educ. Eval.* 55, 68–74. doi:10.1016/j.stueduc.2017.07.003

Buehl, M. M., and Beck, J. S. (2015). "The relationship between teachers' beliefs and teachers' practices," in *International handbook of research on teachers' beliefs*. Editors H. Fives and M. G. Gill (New York, NY: Routledge), 65–84.

Carless, D. (2009). "Learning-oriented assessment: principles, practice, and a project," in *Tertiary assessment and higher education student outcomes: policy, practice, and research*. Editors L. H. Meyer, S. Davidson, H. Anderson, R. Fletcher, P. M. Johnston, and M. Rees (Wellington, New Zealand: Ako Aotearoa & Victoria University of Wellington), 79–90.

Cizek, G. J. (2010). "An introduction to formative assessment: history, characteristics, and challenges," in *Handbook of formative assessment*. Editors H. L. Andrade and G. J. Cizek (Oxon, United Kingdom: Taylor & Francis), 3–17.

Cohen, L., Manion, L., and Morrison, K. (2008). *Research methods in education*. London, United Kingdom: Routledge.

Decreto-Lei 55/2018 de 6 de julho (2018). Diário da República no 129/2018, 1.ª série. Available at: https://data.dre.pt/eli/dec-lei/55/2018/07/06/p/dre/pt/html (Accessed July 06, 2018).

Despacho normativo 1-F/2016 de 5 de abril (2016). Diário da República no 66/2016, 2.ª série. Available at: https://dre.pt/application/conteudo/74059570 (Accessed April 05, 2016).

Dixson, D. D., and Worrell, F. C. (2016). Formative and summative assessment in the classroom. *Theory into Pract.* 55 (2), 153–159. doi:10.1080/00405841.2016.1148989

Eivarsen, K., and Våland, T. I. (2010). "From research question to research design: challenges of obtaining valid sensitive data," in Paper presented at the 26th industrial marketing and purchasing group, Budapest, Hungary (IMP Group). 2–4. September, 2010, 1–13. Available at: http://www.impgroup.org/uploads/papers/7508pdf (Accessed September 11, 2014).

Etikan, I., Musa, S. A., and Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *Am. J. Theor. Appl. Stat.* 5 (1), 1–4. doi:10.11648/j.ajtas.20160501.11

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Fletcher, R. B., Meyer, L. H., Anderson, H., Johnston, P., and Rees, M. (2012). Faculty and students conceptions of assessment in higher education. *High. Educ.* 64 (1), 119–133. doi:10.1007/s10734-011-9484-1

Gipps, C. (1999). Socio-cultural aspects of assessment. *Rev. Res. Educ.* 24, 355–392. doi:10.2307/1167274

Guerra, P. L., and Wubbena, Z. C. (2017). Teacher beliefs and classroom practices cognitive dissonance in high stakes test-influenced environments. *Issues Teach. Educ.* 26 (1), 35–51. Available at: https://files.eric.ed.gov/fulltext/EJ1139327.pdf (Accessed March 9, 2019).

Harris, L. R., and Brown, G. T. L. (2009). The complexity of teachers' conceptions of assessment: tensions between the needs of schools and students. *Assess. Educ. Principles, Policy Pract.* 16 (3), 365–381. doi:10.1080/09695940903319745

Hattie, J. (2012). *Visible learning for teachers maximizing impact on learning*. New York, NY: Routledge.

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

James, M., and Pedder, D. (2006). Beyond method: assessment and learning practices and values. *Curric. J.* 17 (2), 109–138. doi:10.1080/09585170600792712

Karaagac, M. K., and Threlfall, J. (2004). "The tension between teacher beliefs and teacher practice: the impact of the work setting," in Proceedings of the 28th conference of the international group for the psychology of mathematics education. Editors M. J. Høines, A. B. Fuglestad., Bergen, Norway, July 14–18, 2004, Vol. 3, 137–144. Available at: http://emis.impa.br/EMIS/proceedings/PME28/RR/RR276_Karaagac.pdf. (Accessed April 9, 2019).

Miles, M. B., Huberman, A. M., and Saldaña, J. (2014). *Qualitative data analysis. a methods sourcebook*. 3rd Edn. Thousand Oaks, CA: Sage.

Mullis, I. V. S., and Martin, M. O. (2015). *TIMSS 2015. assessment frameworks*. Chestnut Hill, MA: Chestnut HillTIMSS & PIRLS International Study Center.

Nortvedt, G. A., Santos, L., and Pinto, J. (2016). Assessment for learning in Norway and Portugal: the case of primary school mathematics teaching. *Assess. Educ. Principles, Policy Pract.* 23 (3), 377–395. doi:10.1080/0969594x.2015.1108900

Opre, D. (2015). Teachers' conceptions of assessment. *Procedia—Social Behav. Sci.* 209, 229–233. doi:10.1016/j.sbspro.2015.11.222

Organisation for Economic Co-operation and Development, OECD (2008). "Assessment for learning. formative assessment," in Paper presented at the OECD/CERI international conference learning in the 21st century: research, innovation and policy (Paris, France: OECD Publishing). Available at: https://www.oecd.org/site/educeri21st/40600533.pdf. (Accessed April 9, 2019).

Organisation for Economic Co-operation and Development, OECD (2016). *PISA 2015 results (volume I): excellence and equity in education*. Paris, France: OECD Publishing.

Pishghadam, R., Adamson, B., Sadafian, S. S., and Kan, F. L. F. (2014). Conceptions of assessment and teacher burnout. *Assess. Educ. Principles, Policy Pract.* 21 (1), 34–51. doi:10.1080/0969594x.2013.817382

Rahman, M. S. (2017). The advantages and disadvantages of using qualitative and quantitative approaches and methods in language "testing and assessment" research: a literature review. *J. Educ. Learn.* 6 (1), 102–112. doi:10.5539/jel.v6n1p102

Remesal, A. (2006). Los problemas en la evaluación del aprendizaje matemático en la educación obligatoria: perspectiva de profesores y alumnos [Problems in the evaluation of mathematical learning in compulsory education: perspective of teachers and students]. Doctoral dissertation: Universitat de Barcelona, Departament de Psicologia Evolutiva i de l'Educació, Barcelona. Available at: http://hdl.handle.net/10803/2646 (Accessed April 12, 2011).

Remesal, A. (2007). Educational reform and primary and secondary teachers' conceptions of assessment: the Spanish instance, building upon Black and Wiliam (2005). *Curric. J.* 18 (1), 27–38. doi:10.1080/09585170701292133

Remesal, A. (2009). "Spanish students teachers' conceptions of assessment when starting their career," in Paper presented at the symposium perceptions and conceptions of assessment in the classroom: different national perspectives, Amsterdam, Netherlands, August 25–29, 2009 (EARLI), 1–11. Available at: http://www.ub.edu/grintie. (Accessed April 11, 2019).

Remesal, A. (2011). Elementary and secondary teachers' conceptions of assessment: a qualitative study. *Teach. Teach. Educ.* 27 (2), 472–482. doi:10.1016/j.tate.2010.09.017

Santos, L. (2004). "O ensino ea aprendizagem da matemática em Portugal: Um olhar através da avaliação," in Investigación en educación matemática: octavo Simposio de la Sociedad Española de Investigación en Educación Matemática (SEIEM): a Coruña, septiembre 9–11, 2004 (Servizo de Publicacións), 127–154.

Shen, B., McCaughtry, N., Martin, J., Garn, A., Kulik, N., and Fahlman, M. (2015). The relationship between teacher burnout and student motivation. *Br. J. Educ. Psychol.* 85 (4), 519–532. doi:10.1111/bjep.12089

Siarova, H., Sternadel, D., and Masidlauskaite, R. (2017). *Assessment practices for 21 st century learning: review of evidence. NESET II report*. Luxembourg, Europe: Publications Office of the European Union. doi:10.2766/71491

Singh, C. K. S., Lebar, O., Kepol, N., Rahman, R. A., and Mukhtar, K. A. M. (2017). An observation of classroom assessment practices among lecturers in selected Malaysian higher learning institutions. *Malaysian J. Learn. Instr.* 14 (1), 23–61. doi:10.32890/mjli2017.14.1.2. Available at: https://files.eric.ed.gov/fulltext/EJ1150557.pdf. (Accessed October 13, 2019).

Suurtamm, C., Koch, M., and Arden, A. (2010). Teachers' assessment practices in mathematics: classrooms in the context of reform. *Assess. Educ. Principles, Policy Pract.* 17 (4), 399–417. doi:10.1080/0969594x.2010.497469

Vandeyar, S., and Killen, R. (2007). Educators' conceptions and practice of classroom assessment in post-apartheid South Africa. *South Afr. J. Educ.* 27 (1), 101–115. Available at: https://files.eric.ed.gov/fulltext/EJ1150092.pdf. (Accessed April 10, 2019).

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

frontiers in Education

ORIGINAL RESEARCH
published: 09 April 2021
doi: 10.3389/feduc.2021.631857

Leading an Assessment Reform: Ensuring a Whole-School Approach for Decision-Making

Dennis Alonzo[1]*, *Jade Leverett*[2] *and Elisha Obsioma*[2]

[1]University of New South Wales, Kensington, NSW, Australia, [2]Beresford Road Public School, Greystanes, NSW, Australia

The ability of teachers to use assessment data to inform decisions related to learning and teaching defines teaching effectiveness. However, to maximise the benefits of teacher decision-making, there is a need to ensure that all teachers across the school are supported to engage in a whole-school approach to ensuring that all students across different stages are supported. This paper reports on a case study of a school in building an assessment culture with a strong focus on using a range of data for teacher decision-making. We used an auto-ethnography to reflect on our experiences in leading this assessment reform. Using the lens of activity theory, we have identified structural, organisational, social and behavioral factors that contribute to the success of the program.

Keywords: assessment reform, teacher decision-making, use of assessment data, assessment culture, assessment for learning

OPEN ACCESS

Edited by:
Lan Yang,
The Education University of Hong Kong, Hong Kong

Reviewed by:
Jiming Zhou,
Fudan University, China
Lisa M. Abrams,
Virginia Commonwealth University, United States

***Correspondence:**
Dennis Alonzo
d.alonzo@unsw.edu.au

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

Received: *21 November 2020*
Accepted: *15 February 2021*
Published: *09 April 2021*

Citation:
Alonzo D, Leverett J and Obsioma E (2021) Leading an Assessment Reform: Ensuring a Whole-School Approach for Decision-Making. Front. Educ. 6:631857. doi: 10.3389/feduc.2021.631857

INTRODUCTION

Teacher assessment knowledge and skills are critical for improving student learning (Black and Wiliam, 1999; Hattie, 2008). A range of theoretical and empirical evidence support the effectiveness of using assessment to increase student outcomes. However, the effectiveness of using assessment relies on the ability of teachers to constantly adapt their teaching in response to student learning needs and learning development (Mandinach and Gummer, 2016). Teachers do this by using and making sense of different data sources to inform the design of their learning and teaching activities to support individual students, a process that has been proven to increase student learning and engagement (van Gee et al., 2016). The ability of teachers to use assessment data to inform their decisions related to learning and teaching, commonly known as teacher assessment data literacy, defines teaching effectiveness.

Studies on teacher assessment data literacy highlights several issues including a low level of proficiency and self-efficacy (Mandinach and Gummer, 2016), misconception of the process (Kippers et al., 2018) and competing workload demands (Kippers et al., 2018). Despite the importance of data literacy, reforms in schools are often fragmented with teachers feeling that they are not fully supported. To maximise the benefits of teacher decision-making, there is a need to ensure that all teachers within the school are assisted to develop a whole-school approach to ensure that all students across different stages of learning receive the support they need. There is also a problem with of varied understanding of the assessment process, when a common understanding of assessment language and processes is needed for successful implementation of any assessment reform (Davison, 2013).

To address the issues of teacher support and competing understanding of teacher data literacy, this case study will describe a whole-school approach undertaken by one school in Australia to

implement an assessment reform focused on building teacher capacity to develop and implement a school data tracking system to help teachers make informed decisions.

Implementing an Assessment Reform

The roles of principals and other school leaders in implementing educational reforms are undoubtedly one of the biggest factors that enable teachers to effectively implement changes in their practice. The people, processes and tools available to support teachers and students are critical for the successful implementation of the program. This has been highlighted by (Davison, 2013) in the context of A*f*L program. She emphasises that for the successful implementation of A*f*L reforms, program implementers should use the principles of A*f*L to develop, implement, monitor and evaluate the assessment literacy program. Consistent with a view of effective student learning in A*f*L, any reform should consider where individual teachers are in terms of their A*f*L competence, where they need to go, and how best to get there. Hence, any assessment literacy programs aimed at supporting teachers to enhance their assessment competence should begin with evaluating their current level of performance and identifying their training and support needs using a tool that clearly defines the criteria and standards for teacher A*f*L literacy.

A*f*L should be embedded in curriculum and assessment institutionally and pedagogically. Teacher A*f*L literacy programs should start by setting and sharing appropriate learning outcomes, success criteria, and performance standards. These learning outcomes are the teachers' assessment knowledge and skills, the success criteria are the indicators of these knowledge and skills, and the performance standards are the quality of assessment practices that will be used to monitor teacher A*f*L development.

Feedback should be used extensively to provide useful information to individual teachers. As proposed by Davison (2013), one characteristic of a successful assessment reform is "constructive qualitative feedback which helps stakeholders (these include teachers) to recognize the next steps needed for reform and how to take them" (p. 265). The effectiveness of feedback on teachers' performance is supported by studies such as Jensen (2011) in Australia, who found that if teachers receive feedback related to their performance, their effectiveness could rise by up to 30%.

Teacher A*f*L literacy programs should also develop the self and peer assessment capability of teachers (Davison, 2013). Teachers should be encouraged to regularly reflect on their practices to assess how effective they are and how well they are progressing in using A*f*L to improve their professional learning.

A*f*L literacy should provide teachers with continuing opportunities to engage in further education. Contrary to the common practices of most formal training, professional development, AfL programs are most effective if embedded in teachers' everyday classroom activities. Black et al. (2003) emphasise that professional development for teachers to adopt and adapt A*f*L should be framed in such a way that teachers will be fully engaged in a range of activities where they are treated as learners themselves rather than simply telling them how to use assessment and assessment information.

The importance of continuous sharing and reflecting on their practices by teachers and their peers goes far beyond acquiring explicit knowledge. The community of learners they create gives them opportunities to share and acquire tacit knowledge, which cannot be transferred so easily through formal training and conferences. Superficially, it may seem easy to create such a learning environment, but there are a number of critical factors that influence its effectiveness. Amongst these are trust, early involvement, due diligence (Foos et al., 2006), personal interest and shared values (Dhanaraj et al., 2004), intrinsic motivation (Osterloh and Frey, 2000) and fit to the organization (Ambrosini and Billsberry, 2007). It is, therefore, imperative that systems identify and adopt the philosophical changes required for effective assessment A*f*L literacy. Systemic changes should foster trust, develop and communicate shared values, support intrinsic motivation, and find ways for individuals to fit into the school system. The latter requires not only helping teachers to change their assessment practices, but developing personal attributes, which are necessary pre-requisites for A*f*L literacy.

Assessment literacy is not only necessary for teachers but for all other stakeholders, including administrators, students and parents (Davison, 2013). The linkage of assessment literacy to key responsibilities (Popham, 2009) defines its true nature. People with different stakes in education have different needs and so require levels of assessment literacy. Davison (2013), as the lead consultant in assessment reform in Hong Kong, Singapore, and Brunei, found that the most important factor contributing to failure of assessment reform is misconceptions amongst policy makers about what A*f*L really is. Due to lack of understanding of what it takes to implement assessment reform, policy makers may think that simply changing the assessment practices of teachers will make the assessment reform successful. This, however, is not the case because teachers are not autonomous, nor working in isolation. What is required is the establishment of strong A*f*L culture at all levels of stakeholders and across the system (Davison and Leung, 2009).

An effective A*f*L literacy program should recognise the diversity of teachers, who, just like students, have individual needs, diverse learning characteristics and different classroom contexts in which they operate. Hence, A*f*L literacy programs should use the concept of differentiated instruction and adopt various strategies that suit teachers' needs. Above all, program implementers should have a strong belief that all stakeholders can improve their assessment literacy (Davison, 2013).

In summary, the effectiveness of a teacher A*f*L literacy program starts with a clear understanding of the basic principles of A*f*L by stakeholders, re-engineering the educational culture and re-aligning educational practices to A*f*L principles to provide teachers with an environment that models A*f*L culture, providing the necessary support services to teachers, thus enabling teachers to actualize their learning.

Building Teacher Decision-Making

Research on assessment highlights the need to build teacher assessment decisions. For every learning and teaching episode, teachers need to constantly engage in ongoing decisions to

provide the necessary support for individual students. The importance of teacher decision-making was highlighted in the seminal paper by Black and Wiliam (1999) on formative assessment.

> Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited (p. 109).

In that definition, it is evident that gathering and analysing data and using the results to inform the next steps of learning are critical for the effectiveness of assessment to improve student outcomes. Building on the work of Black and Wiliam, the Assessment Reform Group proposed a definition of assessment for learning (A*f*L) as "the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there" (ARG, 2002, p. 2) This definition highlights the need for teachers to constantly gather and use student data to inform all learning processes. This has been clearly articulated in the more recent conceptualisation of teacher assessment for learning literacy that "accounts for knowledge and skills in making highly contextualised, fair, consistent and trustworthy assessment decisions to inform learning and teaching to effectively support both students and teachers' professional learning" (Alonzo, 2016).

Even with the emphasis given in the literature on teacher decision-making, research evidence shows that teachers still have relatively low proficiency in this area (Mandinach and Gummer, 2016). Also, there is more to understanding data for decision-making than simply knowing how to interpret grades, marks and high-stakes tests (Bowers, 2009; Kippers et al., 2018). The low proficiency of teachers and the reliance on standardised student data for decision-making compromises teachers' self-efficacy in decision-making. What is needed in the field is a more practical approach to using a range of data, from "in-class contingent formative assessment to formal summative assessments used for formative purposes" (Davison, 2007), with teacher professional development in the area of A*f*L and decision-making recently given renewed emphasis (Kippers et al., 2018).

From the first conceptualisation of the role of teacher decision-making, there has been a great demand for teacher professional development in assessment decision-making. In a comprehensive list of content for building teacher A*f*L literacy, (Popham, 2009) explicitly listed a number of skills relating to teacher decision-making. Several education bureaucracies have been implementing assessment literacy development programs to train teachers using various modalities including the development of resources and advice. However, despite all these initiatives, the quality of teachers' decision-making remains relatively low.

Building teachers' assessment decision-making is part of a bigger assessment reform that needs to happen in schools. Teacher assessment knowledge and skills must be improved for teachers to make informed decisions. Teachers' A*f*L literacy (Popham, 2011), which include teachers' data literacy (Mandinach and Gummer, 2013) need to be at a certain level of competence. Davison (2013) emphasises that for the successful implementation of assessment reform, program implementers should use the principles of A*f*L to develop, implement, monitor and evaluate the assessment literacy program. From this perspective, it is necessary that teachers should be considered as learners, that is, just like the students they teach they need support to become independent and self-regulated teacher learners. To achieve this, teacher A*f*L literacy programs should be organised around a number of the key principles of A*f*L. Davison (2013) has expanded on seven core A*f*L principles and contextualised them to a teacher A*f*L literacy program, as follows:

1. Consistent with a view of effective student learning in A*f*L, a teacher A*f*L literacy program should consider where individual teachers are in terms of their A*f*L competence, where they need to go, and how best to get there. PD should begin with evaluating their current level of performance and identifying their training and support needs.
2. AfL should be embedded in curriculum and assessment institutionally and pedagogically. Teacher AfL literacy programs should start by setting and sharing appropriate learning outcomes, success criteria, and performance standards. These learning outcomes are the teachers' assessment knowledge and skills, the success criteria are the indicators of these knowledge and skills, and the performance standards are the quality of assessment practices that will be used to monitor teacher AfL development.
3. Feedback should be used extensively to provide useful information to individual teachers. As proposed by Davison (2013), one characteristic of a successful assessment reform is "constructive qualitative feedback which helps stakeholders (these include teachers) to recognize the next steps needed for reform and how to take them" (p. 265). The effectiveness of feedback on teachers' performance is supported by studies such as Jensen (2011) in Australia, who found that if teachers receive feedback related to their performance, their effectiveness could rise by up to 30%. Just as in the classroom, teacher A*f*L literacy learning needs to utilize feedback in order for teachers to have discussions about their performance in relation to the learning outcomes, criteria, and standards.
4. Teacher A*f*L literacy programs should also develop the self and peer assessment capability of teachers (Davison, 2013). Teachers should be encouraged to regularly reflect on their practices to assess how effective they are and how well they are progressing in using A*f*L to improve their professional learning. Similar to student self and peer assessment, teachers need to have criteria and standards readily available to guide them in their reflective practice. Self and peer assessment is only possible if there is an assessment tool that can be used to guide them to assess their performance.

5. A*f*L literacy should provide teachers with continuing opportunities to engage in further education. Contrary to the common practices of most formal training, professional development, A*f*L programs are most effective if embedded in teachers' everyday classroom activities. Black et al. (2003) emphasise that professional development for teachers to adopt and adapt A*f*L should be framed in such a way that teachers will be fully engaged in a range of activities where they are treated as learners themselves rather than simply telling them how to use assessment and assessment information. In other words, teachers should undergo authentic learning that fosters inquiry, experimentation, collaboration, and reflection (James et al., 2007).
6. The importance of continuous sharing and reflecting on their practices by teachers and their peers goes far beyond acquiring explicit knowledge. The community of learners they create gives them opportunities to share and acquire tacit knowledge, which cannot be transferred so easily through formal training and conferences. Superficially, it may seem easy to create such a learning environment, but there are a number of critical factors that influence its effectiveness. Amongst these are trust, early involvement, due diligence (Foos et al., 2006), personal interest and shared values (Dhanaraj et al., 2004), intrinsic motivation (Osterloh and Frey, 2000) and fit to the organization (Ambrosini and Billsberry, 2007). It is, therefore, imperative that systems identify and adopt the philosophical changes required for effective assessment A*f*L literacy. Systemic changes should foster trust, develop and communicate shared values, support intrinsic motivation, and find ways for individuals to fit into the school system.
7. Assessment literacy is not only necessary for teachers but for all other stakeholders, including administrators, students and parents (Davison, 2013). The linkage of assessment literacy to key responsibilities (Popham, 2009) defines its true nature. People with different stakes in education have different needs and so require levels of assessment literacy. Furthermore, Davison argues that issues in A*f*L implementation should be used to develop an assessment literacy program for policy makers. At the highest level, the nature of A*f*L should be clearly understood so as to facilitate the legislation of some pre-requisites needed to institutionalise A*f*L implementation.
8. An effective A*f*L literacy program should recognise the diversity of teachers, who, just like students, have individual needs, diverse learning characteristics and different classroom contexts in which they operate. Hence, A*f*L literacy programs should use the concept of differentiated instruction and adopt various strategies that suit teachers' needs. Above all, program implementers should have a strong belief that all stakeholders can improve their assessment literacy (Davison, 2013).

The Context of the Study

This paper reports on an assessment literacy program in one public primary school in Australia which focused on developing and implementing a school data tracking system to help teachers make informed decisions. The school is part of a wider Learning Community with four other schools within the area, all together comprising 283 teachers and 4,521 students.

The assessment literacy program has been the focus of professional development for the last three years with the goal of building a stronger assessment culture within the school. It aims to build a common understanding of the principles of assessment *for* learning amongst school leaders, teachers, students and parents/carers. Five assessment leaders were designated to work collaboratively with a university partner. Every term, they engage in professional development and then develop an action plan on how to support other teachers with their assessment literacy. Resources and advice are provided.

In the past, teachers at the school participated in a number of professional learning programs on assessment but there was no common understanding nor collective effort to implement effective assessment practices, as there was no clarity or agreement on what assessment practices were effective in supporting student learning. There were various interpretations of what effective assessment looked like and teachers' practices were diverse, but the use of summative assessment was very common. Furthermore, the collection and use of assessment data was done individually by teachers and data sharing across different stages was not evident. Whole school collaboration was missing and whole school student assessment data was not tracked annually to accurately monitor student development throughout their schooling. This meant that individual student assessment data was not accessible to all staff within the school. Student data were kept by individual teachers and were not accessible to teachers taking the same students the following year.

To address these issues, the school became involved in an educational partnership with a university to build teacher capacity in assessment *for* learning (A*f*L) practices and implementation, with the goal to build a strong assessment culture across school. The overall approach of this program was the use of situated learning to connect the new knowledge and skills gained by assessment leaders and teachers to their current responsibilities (Lave and Wenger, 1991; Flores, 2005). This approach is based on a socio-cultural theory of learning that highlights the critical role of peer conversation, negotiation and consensus-making to co-construct knowledge (Lunenberg et al., 2017). To meet the objectives of the program and establish contextual sustainability, five teachers were trained to take leadership roles in assessment. They serve as mentors for their colleagues, facilitate discussions in professional learning groups to identify, discuss and resolve issues and misconceptions related to assessment, and engage collaboratively with their colleagues to develop, implement and evaluate assessment strategies. A teacher A*f*L literacy tool and process are used to help teachers identify their current level of assessment knowledge and skills and key areas of concern. This process helps develop the critical and evaluative skills required to monitor their assessment literacy

(Timperley et al., 2008). Professional learning activities are designed in a way that teachers and school leaders can analyse and evaluate their current level of performance for specific skills against the set of criteria indicated in the tool. From their analysis, individual teachers develop their learning goals for continued engagement in PD activities. To support the suite of professional development activities, the program co-develop resources and templates with evidence-based advice to make the processes accessible for school leaders, teachers, parents and students. These resources have been validated, piloted and evaluated for their effectiveness in supporting the objectives of the program.

METHODS

We use autoethnography to reflect on our experiences as the University Partner (UP), the Instructional Leader (IL) and one of the Assessment Leaders (AL). Autoethnography was chosen as a method because it allows us to draw from our own experience the data needed, analyse it and understand the cultural shift (Campbell, 2016) in assessment practices in one school. It allows a researcher-practitioner to tell their accounts using critical inquiry embedded in theory and practice (McIlveen, 2008).

We adopted the approach of Ellis et al. (2010) by using autoethnography both as a process and a product. As a process, we reviewed program documents, teacher and student data and classroom observation records to guide us in our reflection (Goodall, 2001). These documents facilitated our analysis of our lived experiences about the program. As a product, we reflected on these documents and started to write our own personal account of the program using rich narratives. We focus on what processes, products, engagement and commitment have established the current assessment culture in our school where teachers are strategically designing a range of assessment tasks to elicit student learning and using all these pieces of evidence to make informed decisions to support individual students. To achieve rigour of our reflection, we convened and discussed our own reflection using the activity theory as the theoretical lens to interpret our experiences. The use of a specific theory to interpret our reflection ensured coherence and consistency of the results. More importantly, it allowed us to achieve the aim of this paper. Any insights that we did not agree, we discussed it and deleted those that were problematic. All these processes contributed to the trustworthiness of the results.

Theoretical Framework

Engestrom's (1987) activity theory was chosen as a framework for this study, based on Vygotsky's (1978) conceptualisation of the primacy of culture rather than individual cognition in mediating action, learning and meaning making. In other words, learning is facilitated by social interactions of individuals within the community. This model is useful for understanding how different factors work together to influence various socially and culturally mediated activities to achieve the intended outcomes. It has been extensively applied on learning and development in work pracices.

Specifically, this theory describes the roles of the objects (experiences, knowledge and physical products), tools (documents, resources, etc), and community (people or stakeholders). The subjects, which are the people engaged in the activity, work as part of the community to achieve the object or the outcome of the activity. The quality of the interactions among objects, tools and the community determine the quality of the outcomes. Thus, this analytical framework is useful for reflecting on different elements of social learning systems to understand the patterns of social activities and development, which consequently brings the intended outcomes.

In this paper, the subjects are the teachers who are engaged in professional learning to build their assessment capability particularly on using assessment data to inform learning and teaching. These teachers work within the larger learning community of the other four schools supported by the university which leads the professional development (PD) activities. PD resources including a website are co-developed by the university and assessment leaders of the school.

RESULTS

This section highlights our reflections on our experience and expertise in leading the assessment reform in one school. We focus our reflection on what factors have greatly influenced the development of school assessment culture particularly on using data to inform teachers' decision.

Motivation and Aim of the Program

The need for common understanding of assessment principles, practices and processes of A*f*L in the school provided the impetus for the development of the program. This has been initiated with the commitment of the executive staff members to professional development that engaged all teachers in the network of five schools in the wider learning community. The professional development program focused on building a culture of assessment collaboration through implementing assessment schedules, which originally focused on building teachers' capacity in using learning intentions and success criteria, questioning techniques, self and peer assessment, and teacher feedback. The need to focus on teacher decision-making was brought about by teachers' concerns with what to do with the large amount of assessment data gathered throughout the teaching period. In addition, there were existing system-generated data including the results of national assessment tests teachers were expected to use to inform their assessment decision-making but they were uncertain how to reconcile the two sources of data.

The External Partnerships

The school's need for expert support stimulated the development of the university partnership. Together with the university, the school executive team created a school assessment framework and an assessment schedule for three-year implementation. The professional discussions and collaboration were underpinned by the conceptual framework of A*f*L, building a school culture of

clarity and high expectations for A*f*L implementation. One of the most critical features of the university-school partnership was the designation of the five teachers to be assessment leaders and undertake regular professional development program facilitated by the university partner. We form the team together and the university knowledge from extensive research is used as input for discussing the most appropriate approach and content to address the assessment literacy needs of teachers across the school. The development of the program is based on teacher self-assessment using the teacher A*f*L literacy tool (Alonzo, 2016). This tool was used as it is theoretically and empirically supported and adheres to the A*f*L concept of rubrics with clear criteria and descriptions of five level of performance standards.

After a PD program each term, we develop an action plan which outlines the approach and content for in-school professional development including monitoring individual teachers' implementation of A*f*L. Another feature of the collaboration is working with the other four schools in the learning community. Each year, the five schools join the Learning Community Professional Development Day. Assessment experts from different universities are invited to deliver keynotes, with workshops tailored to the needs of teachers conducted. In addition, the collaboration with other network schools has contributed significantly to building the assessment culture of the school. Within these teams, individual staff were chosen to share their successful assessment strategy for the group. We identify best practices from other schools and incorporate them in our action plan.

To further our collaboration, we have built a community of trust where our classrooms are open for observation. As the University Partner, "*I observe classes of teachers and give feedback at the end of the session. I adopted a dialogic-feedback approach where teachers lead the discussion of their performance and I clarify some of their misconceptions and give them actionable feedback to further improve their practice.*" As an Instructional Leader, "*I find it valuable that our university partner observes a number of teachers to ensure that our practices in our school are adhering to the principles of A*f*L. We have also identified teachers who have best practices and use them as exemplars for other teachers to observe while teaching. During Stage Meetings, we discuss our learnings from observing other teachers.*"

To better develop a strategy for whole-school data collection for decision-making, benchmarking activities were conducted in two schools which are known for their data-driven decision-making initiatives. We looked at the approaches used by each school in terms of data collection, analysis and decision-making. From this experience, we developed a spreadsheet that could help teachers analyse the results of pre and post-tests to calculate the learning gains and effect size. The results of comparing pre and pots-tests are just the starting point for teachers to make decisions for individual students. They have to draw from their professional judgment based on several sources of assessment data including anecdotal records, observation, interviews, self and peer assessment, to validate the results of pre and post-tests.

While this analysis is helpful, there is a need for a more sophisticated assessment tool to link the results to individual students' learning needs. As a result of this, the school has subscribed to an external online assessment provider, Essential Assessment, to reinforce the need of teachers for data. The content of the online assessment program is aligned to the Australian curriculum across stages of schooling. Teachers use it to determine the knowledge and skills of students against the Australian Curriculum achievement standards and use the results along with their classroom assessment data to develop differentiated learning, teaching and assessment activities. Teachers reflect on various data sources and use their professional judgment to identify learning needs and support of individual students.

The Internal Mechanisms

To provide a mechanism for a whole-school approach, a school goal of building an assessment literate school culture was embedded into all teachers' Professional Development Program (PDP) goals. This is a mandatory document for all teachers to set their goals for the whole year. The whole-school assessment framework and assessment schedule was co-developed by all teachers to establish consistency between all staff and co-create school expectation on assessment practices. This process ensured that common understanding of assessment knowledge was shared across the school to establish a culture of trust and value. Assistant principals check that the schedules are implemented, and that assessment data are the focus of stage group meetings and are explicitly used to inform decisions. In addition to ensuring teachers assessments are outlined on a whole school assessment schedule, the school executives created stage-based PDP goals of collecting whole stage data that is accessible to the whole school. This provided transparency for student learning across the whole school. Every student's learning, growth and attainment became a stage-based collective priority. All student results became a critical part of collaborative stage-based planning days.

My role as Instructional Leader (IL) was funded to provide in class professional development to all staff while they are teaching. "*I used my theoretical and practical knowledge of AfL and to flexibly model the various domains of teacher AfL literacy in the classroom. I follow up all lessons with debriefing to ensure collegial discussions focusing on ensuring teachers become the learners while I provide feedback for further improvement.*" *The basic principle I communicate with teachers is to use data to guide teaching to "close the gap" between prior knowledge and the intended learning outcomes. I modeled how data-driven decision-making fits well into the AfL initiative in our school. During this professional development I clearly discussed how the five most common AfL practices can me more effective if driven by student data. I clearly demonstrate how sharing learning outcomes can be effectively used if teachers have a clear record of individual students' prior knowledge. I demonstrate how student assessment data informs the identification of the gap between prior knowledge and learning outcomes, which these gaps will guide teachers to support individual students to set their learning goals. Then, I show how the success criteria scaffold what the learning will look like. I demonstrate also how to differentiate the success criteria based on students' prior knowledge. Further, I emphasise how to give feedback linked to the success criteria and learning goals of*

students. More importantly, I encourage teachers to build the capacity of students to engage in self and peer assessment with the aim for students to gather data related to their learning to monitor their progress. The results of self and peer assessment need to be moderated with reference to the teacher's assessment record. The results of moderation are then used to set future goals of individual students.

I have observed that this method is useful for teachers as it gave them a clear direction on how to implement AfL practices with a strong focus on gathering a range of assessment data and using them to inform every aspect of learning and teaching. In the beginning of the project it was observed that there is a wide disconnect between theory and practice and teachers have the difficulty to implement a coherent learning and teaching activity where each assessment activity supports and builds on from each other. Providing teachers with explicit links on how these practices fit together enhanced their understanding of teaching and delivering AfL practices. They have also understood the range of data that needs are elicited and how to integrate all these different data and make sense of them and use their insights to inform learning and teaching activities to support individual students to learn more effectively.

Based on my experience as an Assessment Leader, there are structural and internal mechanisms that influence the implementation of assessment reform in our school. *The clarity and the connection of the theory and the practical side of AfL practices in the classroom were imperative to its successful implementation. Teachers need to have a clear understanding of the importance of AfL and how this will benefit all students. Differentiated staff professional development is also an important factor ensuring that all staff are supported at their level of understanding and using those staff who were competent as Assessment Leaders in the school provided staff extra support in implementing AfL practices. I organise mentoring sessions with my team to ensure that I could differentiate professional development to each staff member. It was important to build, maintain and sustain momentum around AfL practices throughout the school year to ensure successful implementation.*

As an AL it was important to ensure that my team had clarity around what was expected with AfL. I approached my team members with the idea of collaborative practice to best ensure AfL is used effectively in all classrooms I supervised. Using a simple questionnaire tool, I gained an understanding of what they did and did not know and used that information as a guiding tool to support my classroom teachers. I did this by organising teaching observations of myself using AfL tools in my classroom and having prior and conclusion conversations to discuss what they saw and heard. We set high expectations and it is expected that what we discuss after the lesson observations and theoretical conversations will be implemented. Another mechanism is the collaboration among assessment leaders. We have specific time for discussions around AfL in team meetings, what is working, who among the staff need further specific guidance with and students who may require further differentiation.

Our roles, processes and high expectations support individual teachers. Some teachers require more prompting and support. There are those who did not fully understand the benefits of using data to inform their learning and teaching activities. If we identify any teacher who has misconceptions of AfL, we discuss it in a supportive and respectful environment. Reflecting on the experience of those teachers, they were able to come on board and implement the practices consistently and effectively. I would say that building an assessment culture requires shared responsibility and accountability across all classrooms in the school. It was challenging in the beginning of implementation of the program were many teachers required reminders to consistently apply the AfL practices into their everyday practice particularly the gathering and recording of assessment data.

Adapting the Program

The on-going monitoring of the program allows for its flexibility and growth. As the university partner that delivers the professional development, *I regularly seek feedback from IL and ALs about the perceived needs of teachers and use that to plan the future direction of the program.* To allow for more objective feedback, another researcher was employed to explore the status of the assessment culture in the school. The focus was on students' perception of their engagement in assessment. The data from this engagement were used to inform the future direction of the program.

One example of how we exemplify the use of data to inform the direction of the program is how we use the results of teacher self-reflection using the Teacher AfL tool. The initial results showed that the use of rubrics by teachers was s quite low. During the professional development day, this was extensively discussed to try to identify the contributory factors to this. What appeared to be the problem was the absence of a school-wide common understanding and expectations of quality writing. As the university partner, *I discussed the practices of one school we benchmarked, where they have common rubrics for different writing types and how it is differentiated across different year levels to reflect different levels of proficiency.* As the Instructional Leader, *I facilitated the creation of common rubrics for each type of writing. Together with the teachers, we identified the criteria and established the levels of performance across different year levels.* My role as an Assessment Leader *became easier to communicate with other teachers what rubrics to use and how to use it. Every stage meeting, we reflect on our experience in using these rubrics and provide critical feedback to the IL on how to further improve the rubrics. This is an on-going process and we have found across the time we are using the rubrics that it has improved significantly both its contents and the way we used it across different year levels. It facilitated the consistency of teacher judgment, and we felt that it contributed greatly to the reliability and validity of the data we for our decision-making.*

We have demonstrated above how data gathered from teachers' feedback was used to improve rubrics collaboratively. This allowed teachers to co-create differentiated success criteria for each aspect of writing. It deepened their knowledge of the content and provided better support for students to aim and achieve for higher outcomes. As the Instructional Leader, *I have observed that teachers can now provide specific feedback using the rubrics and can point out to students their specific areas for improvement. The consistency of use of the common rubrics ensure also that students*

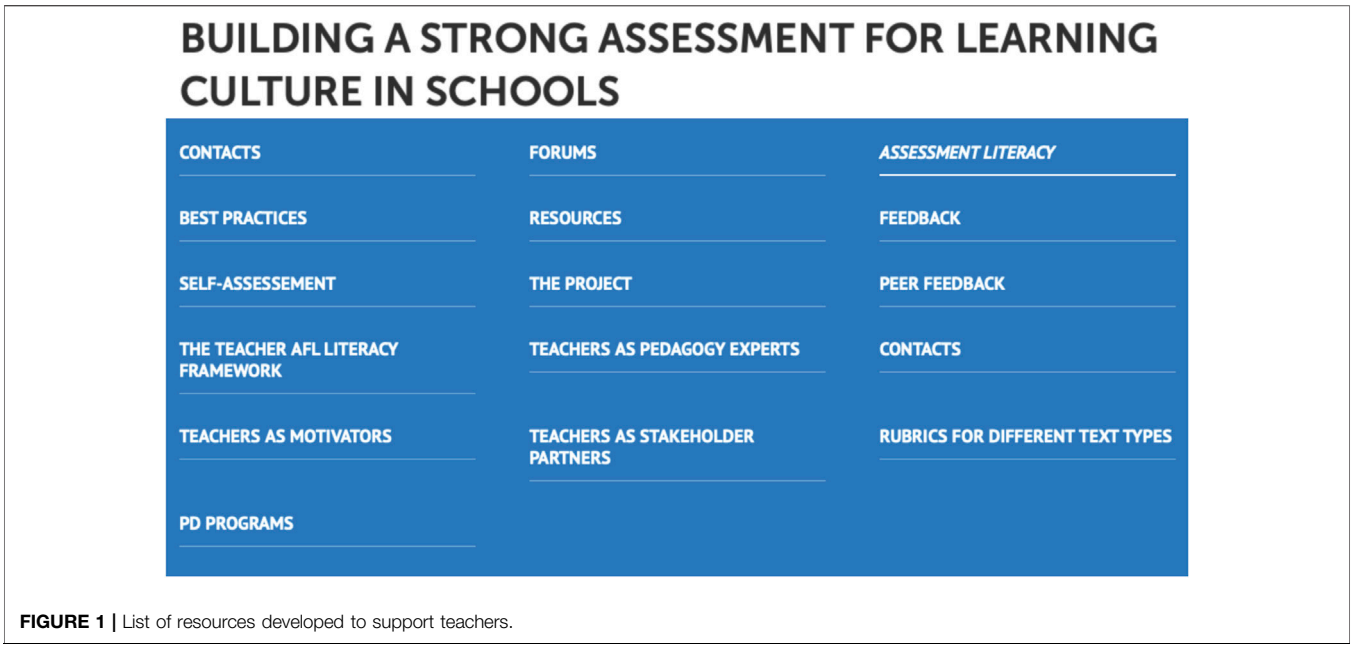


FIGURE 1 | List of resources developed to support teachers.

are assessed based on the same expectations and outcomes. More broadly, the rubrics are used for moderation for consistent teacher judgment. Teachers are fair and develop consistent reporting of student learning and attainment to parents. This process became an integral part of stage meetings.

The process of using rubrics is now well understood by teachers, and the rubrics themselves have become a formative assessment tool for every writing lesson for both the teacher and the students. Teachers use the rubrics to mark the pre-test writing of students and discuss with them their goals based on the results. Teachers record students' pre and post-test student data on a stage-based data tracking sheet. The student growth data became a strong focus of stage-based discussions. As an Assessment Leader. *I have observed also that students' engagement in self and peer assessment has improved significantly as because of their familiarity of the rubrics. They can clearly articulate their learning with reference to the criteria and standards.*

The Resources Needed

Further to professional development, we recognised the need to build resources that will support teachers. We work collaboratively, to identify what resources will be accessible for teachers. Apart from the common rubrics discussed above, we have developed resources aligned to the domains of teachers' A*f*L literacy. We have developed various resources including background information about the program, various A*f*L practices, forums for raising questions and issues, a blog for sharing teachers' best practices and links to various empirical evidence. We put it in a secure website for accessibility and convenience. The frontpage of the website is shown in **Figure 1**.

Observable Outcomes

As a result of the school processes, practices, policy and people leading the assessment reform in the school, teachers became more capable of understanding and using assessment data to guide learning.

This is evident in the discussion during Stage meetings on how teachers refer to student data as their bases for adapting their teaching. The conversation is focused around the clarity of the aims of eliciting and gathering different types of data, making sense of these data and identifying specific actions with the aim to support individual students. We have observed also that teachers now are more confident to use their agency to trial and use more formative assessment practices across the school. Whatever assessment strategy teachers use, it became a school expectation to record observations on formative assessments grids, scaffolds or the teacher's personal record book in all lessons. After two years of implementation, data about students learning has begun to shift the learning and teaching within the classroom. My observation as an Instructional Leader is that, *the more the teachers learn about individual students in terms of their background, learning development and needs, then they are able to provide specific feedback that further scaffolds student learning. The students themselves become teachers of their own learning. They became better at self-assessment and self-directed learners.*

This is where the teacher decision making process has started to impact student learning. When the teacher uses classroom-based assessment data including anecdotal records, they gain a deeper understanding of student needs and are better able to effectively differentiate the school's stage-based teaching programs. Teachers have the knowledge of individual students and are able to flexibly adjust various aspects of learning, teaching and assessment activities to account for individual differences whilst meeting high expectations.

Student goal setting across the whole school has become a new school target. Teachers work closely with students to set their learning goals. Through this process, *students understand what they need to learn and how success looks at each lesson. This develops students' independence, accountability and responsibility into their learning success. It gives them a clear guide to know when*

they are meeting the learning outcomes. We have observed that those students who have clear focus on their goals, engage more on challenging tasks. The conversation with their teachers around their progress using data as evidence enable them to establish their next learning goals. This on-going conversation with individual students enables teachers to see the gaps in student knowledge rather than at the end of the term when post-test is administered. This in-class data gathering gives more data for teachers to use to support students.

Teachers link the goals of the students to the *National Literacy and Numeracy Learning Progressions*. Setting the goals on the Progressions ensured teachers were able to accurately track student attainment of these goals. It also supported whole school consistency in embedding school-based student data in a central and easily accessible location that stays with the student throughout their schooling years.

The assessment data became the focus for engaging parents as well. Parent feedback forms containing learning progress and learning goals are sent home to seek inputs on the learning goals of their children. This is a process that is valued across the whole school community.

DISCUSSION

Our reflection in leading a whole school approach to building an assessment culture with a strong focus on using a range of data for teacher decision-making highlights the different aspects of the program that contribute how it is gaining a significant traction. Based on our reflection, developing teacher decision-making knowledge and skills are influenced by system, organizational structure and interpersonal factors (Marsh and Farrell, 2015; Schildkamp, 2019). These factors are enabled by people leading the reform, processes institutionalised, tools developed and used, and principles adhered to.

Using activity theory, the first requisite for building teachers' decision-making skills is ensuring that everyone in the school has common understanding of the principles of effective assessment practices. For teachers, various contributary factors build their A*f*L literacy, which consequently enable them to engage in data-driven decision making processes starting with eliciting and gathering individual students' learning to making sense of the different types of data and use any insights to inform critical decisions related to improvising student learning and outcomes (Schildkamp, 2019).

The clarity of the aims of the program and consistency of implementation across the school contribute to the development of common understanding among teachers, students and parents. This is a critical phase of the program implementation where all teachers have high level of understanding of not only the aims of the program but also the principles of effective assessment practices (Davison, 2013). The Instructional Leader, Assessment Leaders and teachers are the **subjects** of this initiative. The IL and ALs play critical roles in leading and implementing the assessment plan and they work collaboratively to achieve the aims of the program. They provide the social structure needed for the successful implementation of the program (Poulton, 2020). They serve as mentors to model assessment practices and also monitor the consistency of practice and implementation across school as research shows a lack of leadership in assessment reform contributes to its failure (Marsh and Farrell, 2015).

The **tools** co-developed by these subjects and by the university partner facilitated the development of the common understanding and language of assessment across the school. The availability of the tools is important for supporting teachers to successfully implement their PD assessment goals. For example, the resources provide teachers with materials to deepen their theoretical and empirical knowledge in assessment and decision-making. The common rubrics develop consistency of judgment among teachers, which ensures the trustworthiness of assessment decisions. The co-design process provides the initial link between theory and practice where research evidence and teacher experience are used as inputs for the development of different tools. Closing the gap between theory and practice is an important consideration in assessment reform (Oo, 2020). The effectiveness of the tools depends on their accessibility. For example, the Essential Assessment becomes a handy tool for all teachers that they use anytime they want to check the progress of their students. The design of this assessment tool contributes to its adoption because it lessens teachers time to mark, analyses results and identifies micro skills that individual students have achieved and suggests learning goals. Another key to the effectiveness of these tools lies on creating a culture of continuous improvement of these tools. The opportunity during stage meetings to reflect on what aspects of the tools need to be revised values individual teachers' voice and hence, creating a culture of trust.

The internal mechanisms constituted by the **rules** like the use of common rubrics, stage meetings, embedding of assessment goals to PDP goals, moderation and classroom observations emerged through collective agreement. These are not imposed rules but rather developed through respectful and dialogic conversation with teachers, thus more likely to be acted on by all teachers. They are agreed with a common understanding that these rules will help everyone in achieving the aims of the program. These rules have created clarity of expectations and built positive relationships amongst IL, ALs and teachers which has contributed greatly to the success of the program (Poulton, 2020). Through these rules, everyone becomes responsible and accountable of student learning. Any rules that are counter-intuitive and are not supporting the aim of the program are discussed and modified.

The university partnership and participation in a wider school network's activity created the wider **community** for collaboration, which facilitates translation of theory into practice and sharing of best assessment practices. The collaboration between university partner and the school allow for the critique of research evidence and how it can be applied in the school context. IL, ALs and teachers try some strategies and then later evaluate their effectiveness (a separate paper on university-school partnership is in-preparation to highlight this critical aspect of assessment reform). Drawing from their experience they can verify the theoretical knowledge generated from research in the university. Through this process, research

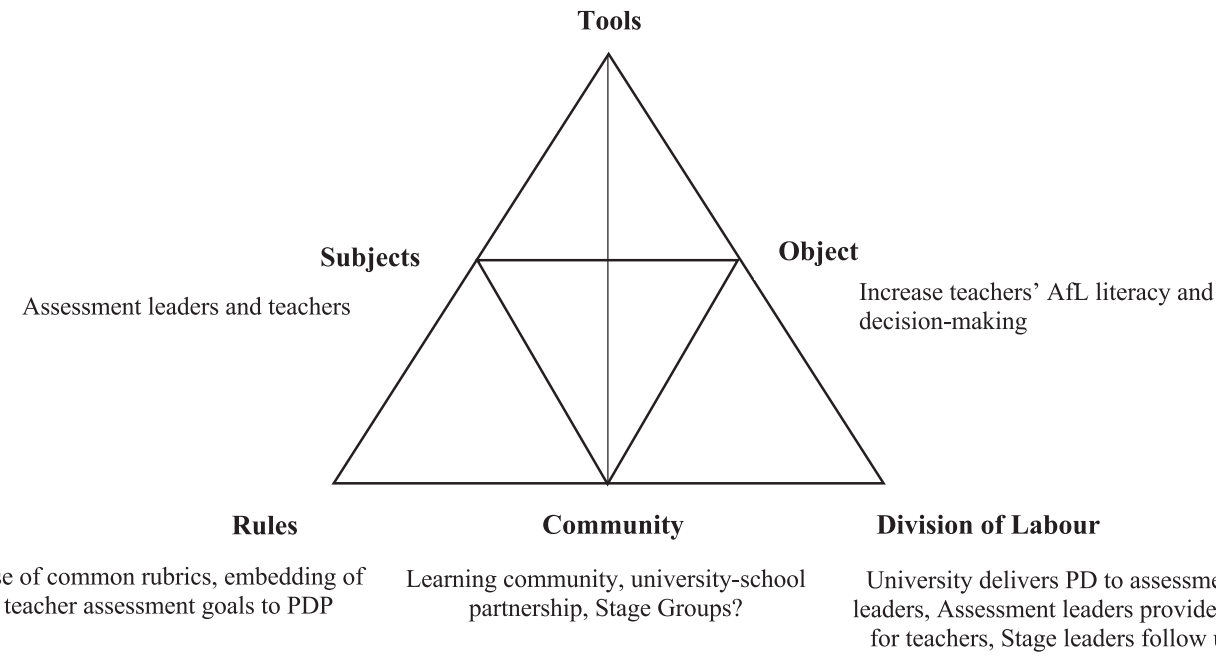


FIGURE 2 | The interrelationships of the factors that contribute to the success of the program.

outputs are translated into practical skills for teachers, which in turn enhances their AfL practices and decision-making. The community and the relationship created provide the environmental factors that facilitate the growth of the program (Marsh and Farrell, 2015).

The participation of different key people supporting the program with the responsibilities clearly articulated is a **division of labour** which provides the critical personnel to implement and evaluate every aspect of the program. The specific roles played by the leadership team including the principal and school executives, the allocation of resources and the funding and appointment of assessment leaders are the school-level factors underpinning its success (Cosner, 2011). The trust of the principal and the clarity of roles of the IL and ALs have contributed to their capacity to take full responsibility of the program. Whilst strong support is provided to teachers, IL and ALs encourage teachers to use their agency to try implement any assessment strategies that they think could help their students based on the data. The degree of flexibility and tapping into teacher agency facilitate changes in teacher practices (Priestley et al., 2012). This flexibility allows for teachers to develop their adaptability in assessment, an important consideration to ensure effective implementation of assessment (Loughland and Alonzo, 2019). In this process, teachers are constantly reflecting on how assessment can be best implemented in different context. This adheres to the context-drive nature of assessment.

All these factors are illustrated in **Figure 2**.

CONCLUSION

Although this paper is based on our collective reflection only, it provides an extensive overview on how to lead a school-wide assessment reform to build a strong assessment culture that can grow teachers' capacity in decision-making. We have demonstrated the function of various tools, the school commitment and internal processes, the partnership created to support the school and the roles of the key players. To further substantiate our claims, however, empirical data is needed to be gathered to provide evidence of the impact of the program on the practices of teachers and how their practices increases student outcomes, which is the focus of our current work.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

DA: Conceptualisation, methodology, analysis and interpretation, writing-reviewing and editing. JL: Conceptualisation, methodology, analysis and interpretation, writing-reviewing and editing. EO: Conceptualisation, analysis and interpretation, writing-reviewing and editing.

REFERENCES

Alonzo, D. (2016). Development and application of a teacher assessment for learning (AfL) literacy tool. Available at: http://unsworks.unsw.edu.au/fapi/datastream/unsworks:38345/SOURCE02?view=true (Accessed July 20, 2020).

Ambrosini, V., and Billsberry, J. (2007). "Person-organisation fit as anamplifier of tacit knowledge," in Paper presented at the 1st global e-conference on fit.

ARG (2002). *Assessment for learning: 10 principles*. London, United Kingdom: Nuffield Foundation.

Black, P., Harrison, C., Lee, C., Marshall, B., and Wiliam, D. (2003). *Assessment for learning: putting it into practice*. Oxford, United Kingdom: Oxford University Press.

Black, P., and Wiliam, D. (1999). *Assessment for learning: beyond the black box*. Cambridge, United Kingdom: University of Cambridge School of Education.

Bowers, A. J. (2009). Reconsidering grades as data for decision making: more than just academic knowledge. *J. Educ. Adm.* 47, 609–629. doi:10.1108/09578230910981080

Campbell, E. (2016). Exploring autoethnography as a method and methodology in legal education research. *Asian J. Leg. Educ.* 3 (1), 95–105. doi:10.1177/2322005815607141

Cosner, S. (2011). Teacher learning, instructional considerations and principal communication: lessons from a longitudinal study of collaborative data use by teachers. *Educ. Manag. Adm. Leadersh.* 39 (5), 568–589. doi:10.1177/1741143211408453

Davison, C. (2013). "Innovation in assessment: common misconceptions and problems," in *Innovation and change in English language education*. Editors K. Hyland and L. Wong (Oxon, United Kingdom: Routledge), 263–275.

Davison, C., and Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Q.* 43 (3), 393–415. doi:10.1002/j.1545-7249.2009.tb00242.x

Davison, C. (2007). Views from the chalkface: English language school based assessment in Hong Kong. *Lang. Assess. Q.* 4 (1), 37–68. doi:10.1080/15434300701348359

Dhanaraj, C., Lyles, M. A., Steensma, H. K., and Tihanyi, L. (2004). Managing tacit knowledge and explicit knowledge transfer in IJVs: the role of relational embeddedness and the impact on performance. *J. Int. Bus. Stud.* 35 (5), 428–442. doi:10.1057/palgrave.jibs.8400098

Ellis, C., Adams, T., and Bochner, A. (2010). Re: forum qualitative sozialforschung/forum: qualitative social research. *Forum Qual. Soc. Res.* 12, 1–8. doi:10.17169/fqs-12.1.1589

Engeström, Y. (1987). *Learning by expanding: An activity-theoretical approach*. Helsinki: Orienta- Konsultit.

Flores, M. A. (2005). How do teachers learn in the workplace? findings from an empircal study carried out in Portugal. *J. In-service Educ.* 31, 485–508. doi:10.1080/13674580500200491

Foos, T., Schum, G., and Rothenberg, S. (2006). Tacit knowledge transfer and the knowledge disconnect. *J. Knowl. Manag.* 10 (1), 6–18. doi:10.1108/13673270610650067

Goodall, B. H. L. (2001). *Writign the new ethnography*. Walnut Creek, CA: AltaMira.

Hattie, J. (2008). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. Hoboken, NJ: Routledge.

James, M., Black, P., Carmichael, P., Drummond, M. J., Fox, A., MacBeath, J., et al. (2007). *Improving learning how to learn: classrooms, schools and networks*. London, United Kingdom: Routledge.

Jensen, B. (2011). Better teacher appraisal and feedback: Improving performance. Retrieved from Grattan Institute, Melbourne: Available at https://grattan.edu.au/wp-content/uploads/2014/04/081_report_teacher_appraisal.pdf

Kippers, W. B., Wolterinck, C. H. D., Schildkamp, K., Poortman, C. L., and Visscher, A. J. (2018). Teachers' views on the use of assessment for learning and data-based decision making in classroom practice. *Teach. Teacher Educ.* 75, 199–213. doi:10.1016/j.tate.2018.06.015

Lave, J., and Wenger, E. (1991). *Situated learning: legitimate peripheral participation*. Cambridge, United Kingdom: Cambridge University Press.

Loughland, T., and Alonzo, D. (2019). Teacher adaptive practices: a key factor in teachers' implementation of assessment for learning. *Aust. J. Teach. Educ.* 44 (7), 18–30. doi:10.14221/ajte.2019v44n7.2

Lunenberg, M., Murray, J., Smith, K., and Vanderlinde, R. (2017). Collaborative teacher educator professional development in Europe: different voices, one goal. *Prof. Dev. Educ.* 43 (4), 556–572. doi:10.1080/19415257.2016.1206032

Mandinach, E. B., and Gummer, E. S. (2013). A systemic view of implementing data literacy in educator preparation. *Educ. Res.* 42 (1), 30–37. doi:10.3102/0013189x12459803

Mandinach, E. B., and Gummer, E. S. (2016). What does it mean for teachers to be data literate: laying out the skills, knowledge, and dispositions. *Teach. Teacher Educ.* 60, 366–376. doi:10.1016/j.tate.2016.07.011

Marsh, J. A., and Farrell, C. C. (2015). How leaders can support teachers with data-driven decision making. *Educ. Manag. Adm. Leadersh.* 43 (2), 269–289. doi:10.1177/1741143214537229

McIlveen, P. (2008). Autoethnography as a method for reflexive research and practice in vocational psychology. *Aust. J. Career Dev.* 17 (2), 13–20. doi:10.1177/103841620801700204

Oo, C. Z. (2020). "Assessment for learning literacy and pre-service teacher education: Perspectives from Myanmar. PhD thesis. Sydney, Australia: School of Education, University of New South Wales.

Osterloh, M., and Frey, B. S. (2000). Motivation, knowledge transfer, and organizational forms. *Organ. Sci.* 11, 538–550. doi:10.1287/orsc.11.5.538.15204

Popham, W. J. (2009). Assessment literacy for teachers: faddish or fundamental? *Theory Pract.* 48 (1), 4–11. doi:10.1080/00405840802577536

Popham, W. J. (2011). Assessment literacy overlooked: a teacher educator's confession. *Teacher Educ.* 46 (4), 265–273. doi:10.1080/08878730.2011.605048

Poulton, P. (2020). Teacher agency in curriculum reform: the role of assessment in enabling and constraining primary teachers' agency. *Curric. Perspect.* 40 (1), 35–48. doi:10.1007/s41297-020-00100-w

Priestley, M., Edwards, R., Priestley, A., and Miller, K. (2012). Teacher agency in curriculum making: agents of change and spaces for manoeuvre. *Curric. Inq.* 42 (2), 191–214. doi:10.1111/j.1467-873X.2012.00588.x

Schildkamp, K. (2019). Data-based decision-making for school improvement: research insights and gaps. *Educ. Res.* 61 (3), 257–273. doi:10.1080/00131881.2019.1625716

Timperley, H., Wilson, A., Barrar, H., and Fung, I. (2008). Teacher professional learning and development: best evidence synthesis on professional learning and development. Available at: https://researchspace.auckland.ac.nz/bitstream/handle/2292/12537/TPLandDBESentire.pdf?sequence=4 (Accessed June 16, 2020).

van Geel, M., Keuning, T., Visscher, A. J., and Fox, J.-P. (2016). Assessing the effects of a school-wide data-based decision-making intervention on student achievement growth in primary schools. *Am. Educ. Res. J.* 53 (2), 360–394. doi:10.3102/0002831216637346

Vygotsky, L. (1978). *Mind and society: The development of higher mental processes*. Cambridge, MA: Harvard University Press.

Conflict of Interest: DA is the editor of this journal. He inhibited in the process of editorial review.

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.

frontiers in Education

ORIGINAL RESEARCH
published: 12 April 2021
doi: 10.3389/feduc.2021.619319

Explicating the Value of Standardized Educational Achievement Data and a Protocol for Collaborative Analysis of This Data

Bronwen Cowie, Frances Edwards and Suzanne Trask*

Division of Education, University of Waikato, Hamilton, New Zealand

Governments expect teachers to be able to make sense of and take action on data at various levels of aggregation. In our research we collaborated with 13 teachers from six primary schools and one intermediate school to use a Data Conversation Protocol to analyze and act on mathematics assessment data generated through a standardized assessment tool—the Progressive Achievement Test (PAT). Our intention was to optimize teacher use of this data for pedagogical decision making and action. At team meetings, the teachers co-constructed then refined a taken-as-shared definition for teacher data literacy for instructional action, which acted to inform and anchor our collaborative research. Data were collected in all teacher meetings and via interviews. Initial findings indicate that a 'Data Conversation Protocol' is helping teachers to slow down the process of considering, interpreting and making a judgement about their students' understanding thereby opening up a space for deeper consideration of the range of possible reasons for student responses to assessment items. Students responded positively to teachers' data informed small group teaching, gaining in understanding and confidence. Teachers considered this confidence translated to more positive engagement with mathematical ideas. Patterns and trends in student responses emerging from the teachers' collaborative analysis of standard data supported a shift from viewing student responses as linked to student or school characteristics to critical analysis of how their teaching approaches might have contributed to student answers/misunderstandings. This finding has implications for how we might challenge assumptions about students through a willingness to engage critically with student achievement data. The importance of teachers having a rich pedagogical content knowledge as a basis for this was clearly evident.

Keywords: data literacy, data conversation protocol, pedagogical decision making and action, standardized data, mathematics

OPEN ACCESS

Edited by:
Dennis Alonzo,
University of New South Wales, Australia

Reviewed by:
Divya Varier,
George Mason University, United States
Lisa Zimmerman,
University of South Africa, South Africa

****Correspondence:***
Bronwen Cowie
bronwen.cowie@waikato.ac.nz

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

Received: *20 October 2020*
Accepted: *15 February 2021*
Published: *12 April 2021*

Citation:
Cowie B, Edwards F and Trask S (2021) Explicating the Value of Standardized Educational Achievement Data and a Protocol for Collaborative Analysis of This Data. Front. Educ. 6:619319. doi: 10.3389/feduc.2021.619319

INTRODUCTION

Day to day teachers in New Zealand, and other jurisdictions that adopt a non-prescriptive or framework approach to curriculum, enjoy considerable agency in matters such as choice of teaching approaches, the detail of program design and how they assess their students. Given this, the basis and nature of teacher decision-making is of crucial importance. Over the last decade the expectations for

teacher use of data as a basis for instructional decision-making have increased (Pierce and Chick, 2011; Schildkamp and Poortman, 2015). Teacher assessment literacy, data-based/data informed decision making and data literacy have emerged as foci for policy and professional development. In this paper our focus is on data literacy. While there is no definitive definition of data literacy it is generally considered to involve teachers establishing a purpose for then collecting, analyzing and interpreting data, and using the insights gained to take instructional action as part of focused inquiry (Datnow and Hubbard, 2015; Gummer and Mandinach 2015; Mandinach and Gummer 2016; Kippers et al., 2018a). This is a complex task, and there is a substantial body of evidence that describes the challenges that teachers face in using data for instructional decision-making and action (e.g., Means et al., 2011; Wayman and Jimerson, 2014; Mandinach and Jimerson, 2016; Schildkamp et al., 2017, ; Visscher, 2020). Moreover, there is evidence that teachers in New Zealand, which is the context of this study, also experience challenges in using data to inform their instructional decision making (Brown and Harris, 2009; Education Review Office ERO, 2018; Edwards and Ogle, 2021; Peter et al., 2017). While intervention research tends to be dominated by studies based in the United States (see Park and Datnow, 2008; Datnow et al., 2012; Marsh, 2012; Athanases et al., 2013), there is evidence of interest elsewhere (Brown and Harris, 2009; Edwards and Ogle, 2021; Kippers et al., 2018a; Kippers et al., 2018b; Lai and McNaughton, 2013a; Lai and McNaughton, 2013b). This work typically positions data literacy as an inquiry sequence similar to that detailed above with data literacy development and enactment relying on a multiplicity of interacting knowledges (statistical, subject content, pedagogical content, curriculum, student, assessment task), teacher mindset and or commitments, and sociocultural/environmental factors (resources, leadership, school culture).

While there has been a sustained emphasis on classroom assessment for formative purposes using teacher generated data (Bell and Cowie, 2001; Black et al., 2003; Ruiz-Primo and Furtak, 2007; Shepard, 2019), changes in technology and increased accountability expectations and measures mean teachers now have access to a wide range of standardized assessment tools and data for classroom use. Research on teacher use of this data is inconclusive, even negative, in terms of its use and impact on classroom level decision making (Stobart, 2008; Lai and Schildkamp, 2013; Volante et al., 2020) suggesting the potential value of this resource is worthy of further investigation.

In this paper we explore teacher consideration and use of data from standardized assessments of student mathematical understanding. Thirteen teachers from seven primary schools and one intermediate school came together to enhance their data literacy skills and explore the instructional potential of data from a widely-used standardized assessment tool. The research project explored the efficacy of a Data Conversation Protocol (DCP) for mediating and supporting processes that embody the principles of productive data analysis, decision-making and instructional action. We illustrate the way the DCP facilitated teacher decision-making and action and conclude that its use permitted teachers to make better founded pedagogical decisions based on root causes rather than symptoms of misconceptions in mathematics. We also detail how the patterns and trends in student responses that emerged from the teachers' collaborative analysis of standard data supported a shift from viewing student responses as linked to student or school characteristics to critical analysis of how their teaching approaches might have contributed to student answers/misunderstandings. These findings have implications for how we might challenge assumptions about students through a willingness to engage critically with student data and support teachers to make greater/more effective use of standardized student achievement data.

SCOPING THE CONCEPTUAL LANDSCAPE FOR TEACHER DATA LITERACY

Three lines of research provide the framing for findings and discussion in this paper. These are:

(1) The definition and importance of teacher data literacy.
(2) Teacher access to and understanding of different kinds of assessment approaches and tools.
(3) The use of standardized data for pedagogical decision-making and instructional action.

Teacher Data Literacy: Definitions, Functions and Practices

Teachers are experiencing increasing pressure from accountability systems, focused on evidence-based teaching and or data-based/informed decision making, with this emphasis designed to address equity and achievement gaps (Means et al., 2011; Klenowski and Wyatt-Smith, 2013; Mandinach and Schildkamp, 2020). Within this agenda the press for teachers to have data literacy skills can be traced to 2001 and the No Child Left Behind initiative in the United States, which emphasized the notion of accountability for student learning outcomes based on standardized test data (Wiener and Hall, 2004). Subsequently, the 2015 Every Student Succeeds Act (ESSA) has provided more flexibility in student achievement tracking but the Act still requires the use of overall accountability measures (U.S. Department of Education, 2015). Currently, a number of states in the United States require school leaders and teachers be evaluated, at least in part, by student achievement data (Ross, 2017). Given this history it is unsurprising that researchers from the United States have been at the forefront of scoping the definition of and practices for data literacy with both the definition and practice still evolving.

Broadly speaking, data literacy can be considered as subsuming, overarching and or distinct from the notion of assessment literacy. Data literacy can be theorized as an individual capacity and one that individuals need to acquire and exercise. It can also be theorized as a collective capacity and set of constantly evolving interconnected practices, grounded in the local context and achieved through collaborative endeavor (Peter et al., 2017). In this paper we view data literacy as a metaconcept (Reeves and Honig, 2015; Cowie and Cooper, 2017;

Beck et al., 2020) with the definition by Mandinach and Gummer (2016) providing the theoretical grounding for our discussion. The Mandinach and Gummer definition is adopted because of its explicit focus on the use of data for instructional action and its expansive view of the kinds of data that can inform this. It also takes account of the breadth of capabilities teachers need to take data-informed instructional action. The definition states:

> *Data literacy for teaching is the ability to transform information into actionable instructional knowledge and practices by collecting, analyzing, and interpreting all types of data (assessment, school climate, behavioral, snapshot, longitudinal, moment-to-moment, etc.) to help determine instructional steps. It combines an understanding of data with standards, disciplinary knowledge and practices, curricular knowledge, pedagogical content knowledge, and an understanding of how children learn (2016, 367)*

While Mandinach and Gummer are concerned with data literacy to inform and enhance instruction there is ample evidence that teachers experience a tension between this agenda and the broader accountability agenda. Brown and colleagues over a number of country contexts (e.g. Brown, 2008; Deneen and Brown, 2016; Brown et al., 2019) have found that while teachers and student teachers consider that assessment can play a productive role in supporting teaching and learning they also view it as having an evaluative function, a negative impact and or as irrelevant. The teachers in the Brown (2008) study identified an improvement focus and a student evaluative function to do with appraising student performance against standards, assigning scores/grades and awarding qualifications. School and teacher evaluative functions were also identified. Irrelevance was associated with rejecting assessment as having a meaningful connection to learning and or believing it to be bad for students. Brown's proposition, and the proposition underpinning the research reported here is that how teachers conceptualize the purpose of assessment and data literacy is important because this influences their practice (Brookhart, 2011; Deneen and Boud, 2014; Barnes et al., 2015; Fulmer et al., 2015). Therefore it is important that interventions focused on developing teacher data literacy for instructional purposes help teachers to reflect on their conceptions of and visions of assessment (Deneen and Brown, 2016). Additionally, a teacher's understanding of the goals and principles underlying a practice are critical because this facilitates the complex, highly situated judgments they need to make without specifying the judgments themselves (Spillane, 2012). Resultant adaptive decisions will always involve, for example, tailoring practice for different groups of students, in specific contexts, as they are engaging with specific kinds of subject matter in order to assist them to achieve valued learning objectives.

Teacher Access to and Understanding of Different Kinds of Assessment Approaches and Tools

There are a plethora of models for the sequence of processes that together lead to data literacy in action/practice, and for how to develop teacher capacity and inclination to work through these. Data literacy interventions typically include an elaboration of the nature of each of the constructs, processes and activities scoped in the Mandinach and Gummer (2016) definition including deciding a focus/goal, data generation methods, data analysis and interpretation, and planning for, taking and reflecting on action, often as an iterative process. Research tends to highlight that teachers can struggle to or may not take instructional action (Kippers et al., 2018a) but there is also evidence that teachers may not have the confidence or knowledge to analyze data in depth (Datnow and Hubbard, 2016; Cowie and Cooper, 2017; Peter et al., 2017; Edwards and Ogle, 2021). Working in the United States, Herman et al. (2015) identified this was the case even for teachers who had access to data from established, high-quality assessments. Van Gasse et al. (2020) and others have identified the tendency to move from data to action with limited consideration of potential causes and or teachers' own assumptions (e.g., Hoover and Abrams, 2013; Jimerson, 2014; Abrams et al., 2015; Bryk et al., 2015; Schildkamp and Poortman, 2015). They recommend paying specific attention to each of the elements of data literacy (see also Bertrand and Marsh, 2015; Farrell and Marsh, 2016). Van Gasse and colleagues point out that each of the elements requires different knowledge and skills. For example, in our study, in order to make decisions about what to focus on based on a summary report of standardized mathematics data for their class, teachers needed to understand how to read data displays; whereas in order to make decisions about how to teach multiplicative thinking (an identified area of weakness), teachers needed to understand multiplication and the range of ways their students might conceptualize multiplication. In this way each aspect of the data use cycle involves a different kind of knowledge and decision about meaning and priority.

Collaboration among teachers where this includes examining student data together is a commonly recommended strategy for developing and supporting teacher data literacy (Love et al., 2008; Hubbard et al., 2014; Bertrand and Marsh, 2015; Reeves and Honig, 2015; Van Gasse et al., 2017; Visscher, 2020). Specifically, professional collaboration around data use is most productive when it is guided by a broader purpose such as providing equitable and excellent education for students (Datnow and Park, 2019; Visscher, 2020). The proposition is that collaboration can help address the challenges individual teachers face in interpreting data, diagnosing problems and formulating action (Gummer and Mandinach, 2015; Datnow and Hubbard, 2016). Through discussion teachers can revisit their initial explanations for poor student results and reflect upon how these results might be linked to their instruction (Bertrand and Marsh, 2015). This said, there is evidence that collaboration is fraught with complexities of power, trust, and diverse priorities (e.g. Daly, 2012). Teacher attitudes towards and motivations for data use along with their self-efficacy and mental models for data use have been identified as influencing their willingness to collaborate (Datnow et al., 2012; Hubbard et al., 2014; Jimerson, 2014; Van Gasse, et al., 2017). It is therefore important to establish a shared understanding of both the instructional action goal of data literacy and the norms for social interaction around data, such as no blame, collective

responsibility, mutual respect (Schildkamp and Poortman, 2015). Teachers need to feel free to take risks and learn from their mistakes in the knowledge they will be supported in the process of experimentation and exploration (Datnow and Park, 2019). Datnow (2020) argues that this kind of professional collaboration is grounded in a mindset of teacher learning, which also provides for emotional support.

A number of studies have identified the value of tools and routines for supporting the development of individual data literacy and of a collaborative culture for data use. For example, Gearhart and Osmundson (2009) identified the value of protocols that embed a clear and specific process for data use and reflection. Others have demonstrated that protocols can support teacher dialogue and data analysis, interpretation, and use within teacher inquiry (Love et al., 2008; Nelson and Slavit, 2008). When practices, tools, and language are shared among teachers, they can much more readily appreciate and learn from one another because they have a common framework for sense-making and goals for participation and learning (Windschitl et al., 2019). In New Zealand, Lai and McNaughton (2013b) demonstrated the value of shared artefacts such as data-interpretation/analysis resources (e.g., PowerPoint slides of graphs and tables that summarized data comparing school achievement data with national data or that displayed relative performances of groups of students which served as templates for schools to use when analyzing their own data) and of the value of schools establishing partnerships with external experts to assist in the development and use of these resources.

The use of Standardized Data for Pedagogical Decision-Making and Instructional Action

To this point we have focused on the data-use cycle as a whole. Here we turn our attention to the nature of data generation as an element that is often taken for granted. Looking beyond education, increasingly people have access to a range of personal health data sourced from wearable technologies/devices. Fors and Pink (2017) argue that the pedagogic importance of personal data lies in 'how they participate in the constitution of new possibilities that enable people to learn about, and configure, their everyday health in new ways" (59). Put another way they suggest that rather than trying to use data to change behavior, people should use it to expand what it is possible to know, do, and imagine. They propose it is more productive for people pursue what possibilities data open up for them to learn and know differently about elements of their lives that they are already familiar with. Connecting this idea to our research, we are interested in how teachers might collaborate around already familiar standardized mathematics achievement data to open up new possibilities for understanding student learning and informing pedagogical decision-making and instructional action. However, studies from a number of country contexts indicate teachers only make limited use of data generated through externally developed tools to inform classroom decision-making (Schildkamp and Kuiper, 2010; Vanlommel et al., 2017; Volante et al., 2020). Volante et al. (2020), based on their review of teacher use of large scale assessments in seven international jurisdictions (United States, Canada, Australia, England, Germany, Finland, and Singapore), suggest three reasons why there is a lack of good formative use of large scale assessments, that is, of national or state-wide compulsory test or examination data. These are: that separate levels of authority promote different use of tests and the data they generate; that large scale tests are often designed and used for accountability purposes so are limited in their scope, and that large scale and classroom assessment remain separated because no-one is advocating for their integration. Other more practical reasons for teachers making limited use of data may be the time elapsed between data collection and data use and issues of curriculum and pedagogical alignment. Additionally, the nature of sampling and complex administration involving multiple tests to different within-class/school participants can mean that it is difficult to disaggregate data derived from large scale tests for use at classroom level. In other words, not all large scale assessments may be fit for purpose in terms of data use at classroom level. On the other hand Anderson (2006) discusses how analysis of question item responses on a national test in Australia can inform pedagogical decision making and action. Pierce and Chick (2011), in their study of Australian Mathematics and English teachers' intentions to engage with externally produced statistical data, found that most teachers considered the data could be used to identify weak students and some teachers (mostly mathematics teachers) thought that they could help to identify curriculum topics that needed attention. In a teacher study in New Zealand, Caldwell and Hawe (2016) concluded that a systematic approach to standardized data was needed for students, teachers, schools and other stakeholders to gain full benefit from the data. A challenge in our research was to know more about the processes and supports needed for teachers to learn with and through data and to exploit any opportunities this might offer to create new and productive opportunities for improving student learning and achievement.

THE RESEARCH CONTEXT AND DESIGN

New Zealand policy documents have consistently emphasized that the primary purpose of assessment is to support learning and teaching (Ministry of Education, 1993; Ministry of Education, 2011; Ministry of Education, 2019). This purpose has consistently been a focus for professional development (Crooks, 2011). Research has emphasized the role of informal on-the-fly and in-the-moment generation of information and action on what students know and can do and might do next (Bell and Cowie, 2001). Planned and more formal assessment has been recognized as having a role to play in interaction with informal and on-the-fly approaches to provide information on whole class and individual student understandings. Classroom based teacher summative assessments (for example, teacher-designed tests and assignments) are recognized as trustworthy and used for reporting and accountability purposes. As noted above, there is recent and increasing interest in teacher data literacy with this being identified as a requirement for high quality assessment practice (Education Review Office ERO, 2018).

The project from which the data for this paper is drawn is a two and a half year government funded Teaching and Learning Research Initiative (TLRI) project (2019–2021) in New Zealand (http://www.tlri.org.nz/tlri-research/research-progress). The TLRI project is using a design-based implementation research (DBIR) approach (Penuel et al., 2011). DBIR research has a focus on persistent problems of practice and a concern with developing theory related to learning and implementation through collaborative design and systematic inquiry, with a longer term aim of developing capacity for sustaining systematic change. The persistent problem of practice that is the research focus is on how to optimize data use for mathematics teaching and learning purposes through a combination of zooming in and out on data at the level of the individual student, class, school and cluster of schools. A second research focus is on TLRI project teachers working as data coaches with their colleagues to develop colleagues' capability to use data for instructional purposes.

In New Zealand teachers can design and or choose what assessment tasks they use, and there is evidence they access a wide range of sources. There are no compulsory national level assessments at the primary school level but the government has made available a number of assessment tools, the New Zealand Ministry of Education supported and New Zealand Council of Education Research developed Progressive Achievement Tests [PATs] being one of these tools. PATs at a range of levels are available for school years 4–10 in reading comprehension and reading vocabulary, in years 3–10 in listening comprehension and in years 3–10 in mathematics. Tests comprise 30–45 multiple choice items depending on the test level. Most questions have alternative conceptions or distractors built in as option choices. The tests are available in paper-based and adaptive online formats. Teachers can access individual question data and class and individual student data reports. Scale scores and stanine information mean a student's level of achievement can be tracked from year to year. For the online version, class individual question response data can be compared with national data (NZCER, n.d.). The project teachers are exploring the potential of PAT mathematics data to inform their instructional decision making and working with colleagues as data coaches to develop their colleagues' capability to use this data. This paper reports on the first year of the project and teacher data-informed action with their classes only.

Teachers from a 16-school Community of Learning | Kāhui Ako (a government funded initiative in which groups of schools in the same area work together to help their students achieve their full potential) were invited to participate in the study with the active consent of their principals. Thirteen teachers from seven different schools volunteered to take part. Ten of the teachers had over ten years teaching experience, the others had taught for between 5 and 10 years. Around a third were the mathematics leader in their school, a third were not, and the remaining third had previously been a mathematics leader in their school.

Four of the schools are full primary schools (Years 0–8), two are contributing schools (Years 0–6) and one is an intermediate school (Years 7–8). Two of the schools were rated within the low end of the socioeconomic ratings in the New Zealand context, two were rated mid-level and three high. The number of students ranged from 118 to 300 for the primary schools. The intermediate school had around 770 students. School student demographics were generally consistent with those New Zealand wide.

In the first year of the project, seven teacher meetings were held, two per term for the first three terms, and one in the final term. At these meetings teachers discussed and then developed a shared definition of data literacy to inform and anchor our collaborative research. Teachers were introduced to a Data Conversation Protocol at the first meeting of the year. They used this to analyze, take action on then report on their class PAT mathematics data.

The Data Conversation Protocol (**Table 1**) was adapted from that developed by Dalton and Anderson (2016). The research team added the "*So then?*" question to ensure teachers were prompted to reflect on the impact of their pedagogical decisions and instructional actions.

Meetings were audio-recorded and field notes taken. Teacher powerpoint presentations on the results of their inquiries were collected as were any materials produced during the meetings. Teachers participated in one-to-one end of year reflective interviews. Interview data was transcribed in full. Audio from teacher meetings was selectively transcribed. Data were collated and analyzed thematically (Braun and Clarke, 2006). Data analysis and findings presentation for this paper reflects the three lines of research that frame this paper with regard to teacher use data to inform practice (consistent with the first research question of the TLRI project).

FINDINGS

In the next section we set out findings related to the study definition of data literacy and its evolution. We also report on the use of a Data Conversation Protocol to support teacher inquiry/analysis of and action on student data. We then outline an example of the impact of these processes on teacher decision making.

Developing, Revisiting and Refining a Definition for Data Literacy

Research emphasises the need for and challenge of developing a shared understanding of goals when teachers undertake research and learning related to data literacy (Jimerson et al., 2020; Mandinach and Schildkamp, 2020). In the TLRI research, teachers and researchers together reviewed available definitions and then co-developed a project definition for data literacy. This was revisited and revised at each meeting. Revisiting and revisioning was deemed necessary because the construct is challenging to define and because it was important that the group had at least a taken-as-shared consensual or compatible understanding (Cobb et al., 1992) of what we were researching together given we were devoting considerable time and effort to the project (Ball et al., 2009; Windschitl et al., 2019). The discussions took place as described below.

TABLE 1 | Data conversation protocol.

Here's what? Describe the data	Describe what you see, just facts, no interpretation or judgement. Mine the data for as much information as possible–look for patterns and probe but stay at the evidence level.	What do you see in the data? What else, specifically? What do you see to indicate that? What evidence can you cite? What patterns do you see? (key trends, common errors, strengths) What might we have missed? Is there other data that would help to understand what is happening?
So what? Interpret the data	Use evidence to seek multiple perspectives and interpretations about what the learner was doing, thinking - what they do/don't understand and can/cannot do? Think about possible causes, assumptions you are making, and evaluate against the data.	Was our assessment fair and valid? What might have been happening here? What evidence suggests this is an option? What might have led to these results and why? What other possibilities might there be? What assumptions are we making here? What don't we know or do we need to find out?
Now what? Implications for teaching	Use evidence and interpretations to raise questions, explore implications for classroom teaching and identify actions to be taken.	What have we learned from our conversation? What question/s does this raise for us? What are some of the implications for our teaching? What is our plan? What are our next steps? What are some of the implications for our assessment for learning actions?
So then? Evidence of student response/learning	Analyze student response for next steps.	Where am I going next? What is the progression of learning I need to consider? What evidence do I need? How/when will I collect it? What do I need to continue to work on with the students? Who still needs support?

What Counts as Data?

The first meeting began with a focus on what counted as data. This was stimulated by the question: When you think about 'data' what comes to mind? Data were described as "numbers, information, trends, results, graphs," but also as "more than just numbers." Teachers questioned whether the data they had access to were always a true reflection of a child's understanding and progress. They discussed whether they would 'trust' data, depending on when and how it was collected and analyzed. An additional prompt was: What sort of 'data' did you use when you made an overall teacher judgement (OTJ). [Until 2010 teachers were expected to collate a range of data to make an OTJ about their students' mathematics achievement relative to a set of nationally mandated standards and this language and thinking has persisted since the requirement was revoked in 2017].

Working Towards a Shared Definition of Data Literacy

The next prompt asked teachers to consider: To you, what is data literacy? Groups discussed and contributed definitions. Examples included: "Data literacy is being able to read and understand data, look for trends and patterns, look for validity and critique, then use data effectively"; "Data literacy is to be able to collect, understand, reproduce and utilize worthwhile data", and "To gather, analyze, act on and reflect on these actions related to data".

Two definitions were then shared:

> *A process that integrates the analysis of educational data to support decisions intended to improve teaching and learning at the school and classroom levels. (Means et al., 2010).*

> *Data literacy is the ability to understand and use data effectively to inform decisions. It is composed of a specific skill set and knowledge base that enables educators to transform data into information and ultimately into actionable knowledge. (Mandinach and Gummer, 2013).*

Teachers were asked to discuss if and how the definitions were consistent with their own ideas. They considered they were. The group then came together to negotiate the definition of data literacy they would use to inform their collaboration in the project. This definition was crafted by recording and adjusting dictated sentences on a whiteboard. The following statement was agreed at the end of this meeting:

> *Data literacy involves collecting/gathering data, analyzing and understanding it and then using this understanding to take action. It includes the knowledge needed to decide if data is worthwhile and or valid and the ability to share information to different groups (Children, other teachers, principal, Boards of Trustees [School governance board] etc.)*

This was the taken-as-shared definition for teacher data literacy. The inclusion of an explicit mention of taking action

on data echoed teachers initial ideas and the focus of the definitions shared with them. Teacher individual interviews indicated that the project commitment to action had been important in them volunteering to be part of the project. The second sentence was a source of more debate. The teachers designed it to encapsulate their concern to maintain a critical stance towards data and their view that information on student learning was of interest and value to a range of decision-makers. This focus was exemplified in the comment, "Parents are interested in their children's progress and we need to be able to communicate this to them". This focus is in line with New Zealand assessment policy (Ministry of Education, 2007; Ministry of Education, 2011).

Revisiting and Refining the Shared Definition

In the second meeting on 24 May, the teachers were reminded of the co-constructed definition and encouraged to refine it. In this session they were talking about data literacy in the context of their own work, and critiquing each other's thinking and the assessment tools they used themselves. For instance, as teachers discussed key issues that examination of PAT data revealed, they discussed the interaction of validity and assessment task design, asking, "What knowledge is needed to unpack a question?" Again, they noted the need to critique both data and the questions that led to it. Commenting, on some of the PAT questions they pointed out that "students can't answer a question involving time differences on an analogue clock if they can't read an analogue clock" and "it is difficult to estimate volume if students are unsure of context or whether to use milliliters or liters".

A Broader View of Data and Data Literacy

During the third workshop on 19 June the team again revisited their data literacy definition. The definition was on display and edited publicly with teachers offering suggestions for additions and refinements. On this occasion the discussion focused on what counted as data in concert with exploring the implications of a holistic view of students and of teachers' responsibility towards their students as 'whole people'. That is, the discussion encompassed the need for teachers to understand how, when and why different data was produced and raised questions about what could and should be considered as relevant for data and data generation when a teacher's goal is to assist students in their learning. The holistic vision of the student as a learner that the teachers endorsed is consistent with current policy within New Zealand (Ministry of Education, 2007). It is also in line with international policy, for example the Every Student Succeeds Act (2015). The teachers' recognition of the need to draw on different types of data and to have an expansive focus when the student group is diverse is supported by research (Gipps and Murphy, 1994; Stobart, 2008; Bernhardt, 2018) and is congruent with system, school, teacher and family interest in the behavioral, affective, and cognitive dimensions of learning and being a learner. As part of their discussion the teachers again raised and discussed the validity of the assessment data they might use: "Had they [students] ever been through this type of question, with this wording? Testing that is too hard provides poor data as when students find the test too hard they give up". They concluded that they needed to interrogate assessment questions *and* student responses. Other points for consideration were whether the questions actually tested what students need to know. The teachers concluded that students needed to have an opportunity to show that they knew and that perhaps this would mean changing some questions.

After several iterations the discussion converged on the following description for data:

> *For us 'data' is a wide range of information including student learning conversations, perceptions, observations, and products of learning, school processes, student demographics (after Bernhardt, 2018) and includes different levels of aggregation.*

Teacher revisiting of the definitions for data literacy has been ongoing, and in 2020 dimensions of vision and data literacy culture have been added to the definition by the teachers. These are the focus of another paper. The taken-as-shared definition for data and for teacher data literacy for instructional action have provided common reference points for teachers when they collaboratively work and talk together. The definitions provide a framework within which the teachers are comfortable to work.

The use of the Data Conversation Protocol: Finding New Possibilities in Familiar (Standardized) Data

The research team introduced a Data Collection Protocol (DCP) to the teachers at the first project meeting in anticipation it would lead to a taken-as-shared way of talking about how they might work with data. The first step of the DCP above prompts teachers to reflect systematically on the data they have generated using the 'Here's what?' prompt. As noted above, teacher focus in the first year of the study was on PAT mathematics data. The mathematics PAT assessments include questions on number knowledge, number strategies, algebra, geometry, and measurement and statistics. The teachers brought their class PAT data to the first meeting. This included individual student responses for each question and test totals with associated scale scores. This information was in tabular form for individuals and scatter plots for classes. No statistical analysis was included. The teachers also had access to national item and item option response distributions for questions for comparison.

The 'Here's what' prompt stimulated a robust discussion on 'What counts as data and for what purposes?' We were all surprised at the level of debate this question generated as discussion probed matters to do with validity, reliability, equity and consequences. Teachers listed and critiqued the nature and potential meanings of the results of commercial tests, of teacher generated tasks, of classroom-based observations and dialogue as well as data on attendance and student mobility across schools. They identified gaps and variations in their individual knowledge in terms of different

assessment tools and knowledge of what data analysis support was available with tools such as PAT. The group then discussed and negotiated a meaning for the other three prompts. Following this, teachers formed small groups and analyzed the class PAT data they had brought to the meeting by looking for patterns within and across their classes and nationally for individual PAT item results. This process included thinking about what actions they could/would take; the "Now what?".

Teachers were then charged to select and take action with 2–5 'target' students. The group decided these students would be selected on the basis that data identified a common misconception or a need for strategy development. For example, one teacher selected target students who had incorrect responses to approximately two thirds of the measurement/geometry questions in PAT Mathematics Test 4. At the next meeting the teachers reported back on their actions with their target students.

Explicating Key Insights From Using the Data Conversation Protocol: Data Analysis

In what follows we outline key themes and insights that arose during the two first year meeting discussions based on the DCP. In essence, the key insights that teachers discussed were the value of taking time to analyze data, of questioning data as evidence of student learning, of considering possible underlying/conceptual reasons for student responses, and of planning for follow up action including revisiting concepts to develop student understanding and confidence.

We have already noted that the teachers critiqued the data they generated through their own classroom based assessment/OTJs and the data generated by commercial sources such as PAT. They concluded that ultimately, "The data is the starting point but we have to know the students" with this point flowing naturally into the "So what?" step of the DCP. Consideration of this step 'slowed down' the teachers' analysis and interpretation of the data they had in front of them. It included consideration of how previous experiences and understanding might be implicated in student responses as in the following teacher comment: "There is a group of students who are flatlining—why are they like that? This can mean going back to previous data to find gaps in understanding." Through comments such as these teachers can be seen to be questioning the data as robust evidence of student learning. The group public consensus was that it was important to not make assumptions about what students did know and could do and what they did not know and could not do nor to assume that skills were in place when they might not be on the basis of one data source. Here teacher data interpretation and critique is in line with Raffe et al. (2019) assertion that, "targeting outcomes without understanding the context or procedural mechanisms that produce them yields constrained insight into how to support and enhance teachers' data use practices." (94). It also suggests that standardized data offers a snapshot of learning and should not be used in isolation or considered as the sole basis to judge learner achievement.

The teachers commented that prior to the TLRI project they might not have analyzed their class PAT data. It was seen as having relevance only for their principal. The following teacher comment is representative of the group view: "Before I would have just marked the PATs, got a stanine to give to my Principal and gone, 'OK'". The teacher continued, "But this time I actually looked at it and thought, 'Oh, measurement...still work to do.' And it has informed my teaching for this term—we're going to go back and revisit measurement this term before they move on next year."

Teacher analysis of possible reasons for student responses to particular questions involved them in taking the time to consider what might be the basis for student answers and to plan for teaching. One teacher noted: "Often you look at the data and we just say they [students] need to work on their addition and subtraction strategies without really narrowing down to look at what do *we* actually need to work on." A representative comment was: "I've never delved as deeply into it [data] before." We can see here that the teachers thought that taking time allowed them to analyze the data more deeply *and* to think about the implications of the teaching approaches they used: "We're all good teachers - but this has given me the time to think about my teaching." The teachers came to the view there was value in taking "a little bit extra time to get proportionately more value". This is an important realisation given evidence from elsewhere that teachers tend to spend very little time and do not consistently analyze student work in depth (Herman et al., 2015). Also important was that slowing down in this way interrupted teacher habits to do with, "This is how I teach this". Careful analysis at the "*So what?*" step meant for one teacher that, "I can actually hone in on students' [ideas and or misconceptions] and ensure planning in relation to that." Through thorough analysis teachers, "Could pin down the issue - and pin down what next". As another teacher explained, "Careful analysis of one question from the PAT test allowed me to really focus my planning and my teaching on the (concept)". Teachers were emphatic that their focused work with small groups of selected students, the "*Now what?*" element of the DCP, was better targeted and more productive. One teacher commented, "Without taking the time to look closely at this data I may have spent less time teaching (the concept) to these students (because I was) assuming that it was more of a calculation error rather than a lack of knowledge."

Explicating Key Insights From Using the Data Conversation Protocol: Action on Data

Typically, teachers' deeper analysis and follow up actions led to their revisiting earlier ideas. In the words of one teacher "I had to go right back but going back helped students move forward." Student responses alerted teachers to the idea that, "We need to take time to cement learning before moving on." They came to appreciate, in line with a number of studies (in the New Zealand context: Alton-Lee, 2003; Nuthall, 2007) that students benefit from encountering ideas multiple times and in multiple contexts: "The process highlighted for me the importance of breadth and providing students with multiple opportunities to grasp a

concept, and more time than I would have previously allowed for." This said, the teachers were clear they did not want their teaching responses to become "gap-filling" even though this might be needed for students to make progress. Datnow and Park (2019) pointed out that teachers need to plan for student growth by identifying student strengths. Comments such as "Celebrating successes and making sure each child has a success to celebrate" indicated that teachers also considered this was important.

In commenting on student responses to their intervention actions the teachers focused on the development of student understanding and the benefit to student confidence that came with understanding an idea that had previously been confusing. One teacher explained this impact as follows:

> *Children can see success—they've figured out one small thing but then that might be one tiny brick in the foundation, and then it snowballs. The kids are keen cos they can do it, instead of a great big concept, they can do one tiny thing then another tiny thing, and so on.*

Other observations were that students now had the confidence to articulate their thinking, were happier in their work and more willing to attempt more complex problems.

A number of teachers commented on how their focused follow up actions with a small group of students had alerted them to the subtle variations in student understanding, or as one teacher explained the situation, "every student is at a different point in their journey." Another teacher elaborated on this point saying, "If I had taught them in a bigger group, they [students] may have missed it, and I as a teacher would have missed subtleties as well." One of the teachers summarized the overall impact of the data inquiry on her own practice and on student learning as follows:

> *I found I was listening more to my students, really paying attention to what their needs were as a targeted group and as individuals. The growth and increased confidence all of these students showed was tremendous.*

The teachers considered their attention to the 'So then' aspect was particularly important in shifting student achievement as it prompted them to review the impact of their actions. One teacher summed up the implications of this focus as: "This is, in some ways, the most important step, especially if the data shows that there is still an issue."

Although the teachers were focusing on data from their own classes, they recognized there were many areas of common concern when they shared ideas across the group. The influence of this collective sharing and recognition of common concerns is explained further in the next section.

An Example of Data Use in Action

Teachers used the Data Conversation Protocol to think about the data as individuals and in small groups then shared ideas with the whole group. During this sharing process eight of 13 teachers identified two-digit subtraction as problematic for their students. Collective analysis and sharing of student choice of answers revealed the commonly accepted answer was a deliberately designed distractor which fitted with students decomposing both numbers and subtracting the smaller 'ones' digit from the larger. One teacher explained: "Say if it was 52–38, they could do 50–30 but then they would just automatically swap the ones digits around because they couldn't do 2–8. So, they just automatically went 8–2 = 6." Important to the subsequent discussion, this pattern of choice was consistent across schools and the years for which the PAT question applied. One teacher explained the impact of sharing with the TLRI group: "Someone brought [the subtraction issue] up and it was, 'Oh, that's right, we have that problem as well.' I thought it was just an issue at our school." Another described discovering, "We all had the same issue of [students] swapping around the ones number. We've just looked at this because of the group, because of the coaching that we've been getting and noticing that in the [geographical area/ Kahui Ako] that there's an issue." The following comment is representative of those from teachers who focused on this issue with a small group of students:

> *Now that I've seen what can happen I'd definitely go with only decomposing the second number, that has helped immensely those kids that were really struggling.*

Another Explained:

> *. . . The Year 7 and 8 students I was working with, once they 'got' what I was telling them, they were just so AMAZED, at what they could do, because they were kids who have struggled all their way through, not achieving where they need to be, and they just looked and went "OHHHHHH! OK!"*

This realisation led to another teacher asking: "Do we set students up correctly for subtraction when we teach addition?" In the context of two-digit subtraction this question related to the commonplace process of teaching two-digit addition as a process of decomposing both numbers and adding. Reflecting on this strategy the teachers' analysis led to their recognition that students' difficulties could be attributed to what (Ryan and Williams, 2007, 23; see also Anderson, 2009) term "intelligent overgeneralization". Ryan and Williams describe this as the tendency to create inappropriate rules based on past experiences, that is to overgeneralize a strategy or rule of thumb. It is of note that all 13 teachers were interested in and confident in teaching mathematics but clearly many had not encountered or thought about the longer-term implications of the decomposition strategy for two-digit subtraction. The group concluded that while the decomposition teaching strategy might be helpful in the short term, they needed to reconsider its use—in the future they would teach students to decompose only the subtractend. From this example the conversation turned to wider consideration of the longer-term consequences of pedagogical strategies with one teacher asking directly: "How does what is being taught at younger year groups/lower levels impact what is taught in subsequent

levels?" Suggestions for common generalizations included: comprehension of the equals symbol, where students understand the equals symbol as "find an answer" rather than "the same as"; multiplication makes bigger; division makes smaller; and longer numbers are always greater in value.

The consistency of student responses across the schools was pivotal in prompting the group to speculate that it might be their teaching strategies and not student or school attributes that were the likely reason for students' answers. In the words of one of the teachers:

> *It was definitely the sharing, just bringing up that issue and then everyone going "Oh yeah, we have that issue". You sort of just think it's our kids . . .. And it was like a catalyst to think what else are we doing that might not just be them. It's definitely something I'm going to share with the rest of the staff as well.*

Teacher discussions, which pooled data and insights from teachers from different schools and school year levels, could be seen to identify and explore what Ball (1993) refers to as "horizon knowledge". Horizon knowledge includes knowledge of how mathematical topics are related over the span of the curriculum. It includes the content and pedagogical content knowledge teachers need to understand the significance of 'what comes before and after' in connection to mathematical ideas. Ball argues knowledge of the mathematical horizon is important because of the role it plays in teacher decision making and because a teacher's choices can anticipate or undermine later development, or what one teacher in our study described as "setting up misconceptions for the future". Teachers in the TLRI project clearly came to appreciate the need to consider this possibility as a consequence of their collaborative analysis of the same standardized assessment data and the DCP.

CONCLUDING COMMENTS

Teacher data literacy and its development are a focus internationally, and in New Zealand which is the context for this paper. The proposition is that data use can inform and enhance teacher pedagogical decision making and action. While the focus is often on teacher formative use of data generated informally through interaction the rise in provision of commercially produced standardized assessments has opened up new and different opportunities for teacher access to data on their students' learning. In this paper we report findings from a study exploring if and how the use of data from a commonplace assessment tool (the PAT) could be used in teacher pedagogical decision-making. PAT assessments have been in use in New Zealand classrooms for over 40 years. Generally, teachers administer the tests and the principal and school leaders' access, analyze, reflect and act on results. Classroom teachers might take a cursory look at their class and individual student results but on the whole the data is not viewed as having direct pedagogical value.

Data Literacy

There is considerable evidence that teachers can struggle to appreciate the pedagogical purpose of data literacy and also that, as a group, teachers can struggle to build a common understanding of the knowledge, habits of mind and language involved in data literacy (Means et al., 2011; Lai and Schildkamp, 2013; Kippers et al., 2018a; Henderson and Corry, 2020). The intervention in the study reported here began by establishing a shared understanding of the nature of and purpose for teachers working together to develop their data literacy. Teachers' critical reflection on and refinement of their co-constructed definition for data literacy each time they met together appeared to be important in sustaining their commitment to and collective ownership of it as a process focused on informing pedagogical decision-making and action (Brown, 2008; Datnow and Park, 2019). The reiteration of its instructional purpose was important in locating their work as counter to their experience that often PAT data was only used by their principals for reporting and accountability purposes. The evolving definition provided a concrete and meaningful anchor and language for their collaborative discussion through its articulation of the process and the purpose for data use—to enhance instruction in support of student learning.

Operationalizing Data Literacy

To operationalise data literacy, we employed a Data Conversation Protocol. This guided teachers in their deeper consideration of what the distribution and detail of their students' PAT results could tell them about their students' thinking and their own practice. Although teacher in depth interpretation of and planning for individual student learning took time, the teachers were convinced that this time was well spent—they were more than pleasantly surprised by their students' responses. As others have found (Datnow and Park, 2019), the teachers were emphatic that they benefited from sharing their experiences with colleagues—student responses and collegial sharing and feedback validated the processes that had been undertaken as worth employing and sharing more widely (teachers are working through a process for this). Teachers using and discussing the question prompts in the Protocol focused teacher attention on the demands of particular assessment items and the patterns of student responses within and across items. Being able to consider these patterns across school years and schools appears to be particularly productive in stimulating the sharing and critical analysis of teaching approaches rather than student attributes. Through their cross school and school year level discussions teachers raised and illustrated the need to consider the possibility of unforeseen consequences of their pedagogical decisions, with the example given in this paper being the longer-term implications of a particular approach to solving addition and subtraction problems. It was the use of a standardized assessment tool (PAT) that allowed the teachers to genuinely share and discuss commonalities in student responses which then impelled them to look beyond individual student attributes and school stereotyping to consider possible implications of their pedagogical approaches. This in turn opened up different foci and options for instructional action. To us, this process has echoes of what Fors and Pink (2017) advocate in relation to the potential reconsideration of familiar data. They propose this

can lead to "the constitution of *new possibilities*" (59), in our study this led to a critical evaluation of a commonplace teaching strategy.

In considering teachers' action on data, Claudet (2020) argument that effective learning interventions need to address underlying *"root causes"* rather than surface-level *"symptoms"* is pertinent. Surface-level symptoms are generally more easily discernible than root causes but if root causes are not identified, then the time teachers spend on symptoms may have limited long term impact. However, identifying root causes takes time and thought. As the teachers reported, identification and action on root causes relies on teachers having in depth content and pedagogical content knowledge (Shulman, 1987). Both were needed for them to interpret the thinking that might underpin student responses. Teachers taking the time to work with a small group of students on a very specific mathematics idea alerted them to the nuances and variations in student thinking and, in some instances, challenged their assumptions about student thinking. Their comments indicated that this experience might have sensitized them to the value of careful analysis and listening going forward. In considering the efficacy of teachers' actions it is also of note that the teachers identified that their students responded very positively to finally making sense of/understanding an idea *and* the confidence students gained from this success translated into their confidence in approaching other ideas/challenges. This further highlights the benefit that might be gained from such small and focused actions.

The Use of Standardized Data

Teachers do not always consider that standardized data generated via externally produced tasks has value for pedagogical decision-making and action (Volante et al., 2020). The teachers and research team chose to focus on standardized PAT data as an opportunity to make greater formative use of data teachers were already obliged to collect. The research shows that standardized data can provide teachers with useful insights into their students' learning, especially when they take time for careful collaborative analysis, as discussed above. Standardized tools are often online and produce a range of pre-designed reports that have the potential to inform and fast-forward teacher decision-making. Teachers in our TRLI study benefited from the range of reports that could be generated from PAT data–item, individual student, class and school. Their analysis and action on the PAT assessment data benefited from the inclusion and detailing of the distractors that were included as options in most test questions. These were based on student alternative conceptions and provided teachers with information about student ideas that they could use to inform their instructional actions (Anderson, 2009; Gierl et al., 2017). In line with the literature the teachers did raise the matter of pedagogical alignment—would their students recognize the question context and format—but they circumvented the matter of timing and curriculum alignment by focusing on particular ideas with small groups of students. In this way they were able to support targeted to students who were/were likely to be struggling with specific and important ideas. As they commented, this was both time consuming and worthwhile.

Looking Forward

Looking forward, it is significant that through their analysis and sharing of standardized data the teachers in this study identified shifts in their focus from students or their own school as being the cause of a learning deficit to consideration of the longer-term impacts of the teaching approaches they were using. Cross school and cross school level sharing using the Data Conversation Protocol was important in this because it prompted teachers to slow down and carefully consider the patterns within and across the data they each had. This challenged rather than confirmed their assumptions (Datnow and Park, 2018) opening up space for new ways of thinking and acting as identified by Fors and Pink (2017), something that is important when then the goal is to enhance instruction for all, and not just some, students. The Data Conversation Protocol and the practices associated with it were both important because together they provided teachers with the agency and tools for better informed decision-making and action. The protocol also provided a basis from which teachers could begin to coach colleagues (a paper in preparation describes the work of teachers as coaches). The project has reported these findings to all sixteen schools in the Kāhui Ako. With principal support we are now working with the TLRI teachers to develop ways to share these insights with teachers in other Kāhui Ako schools.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because Data is confidential to the research team. Requests to access the datasets should be directed to bcowie@waikato.ac.nz.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Division of Education, University of Waikato. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

We have all contributed to this manuscript.

FUNDING

The study that underpins this paper was funded by the New Zealand Ministry of Education Teaching and Learning Research Initiative fund which is administered by the New Zealand Council of Educational Research.

ACKNOWLEDGMENTS

We acknowledge and thank the schools and teachers from the Pukekohe Kāhui Ako (Community of Learners schools) for their participation, in particular Nicola Gibson, Pukekohe Intermediate School who has provided active and insightful leadership to the project.

REFERENCES

Abrams, L. M., McMillan, J. H., and Wetzel, A. P. (2015). Implementing benchmark testing for formative purposes: Teacher voices about what works. *Educ. Asse Eval. Acc.* 27, 347–375. doi:10.1007/s11092-015-9214-9

Alton-Lee, A. (2003). *Quality teaching for diverse students in schooling: Best evidence Synthesis iteration [BES]. Report from the medium term strategy policy division*. Wellington: Ministry of Education.

Anderson, J. (2009). Using NAPLAN items to develop students' thinking skills and build confidence. *Aust. Math. Teach.* 65 (4), 17–23.

Athanases, S. Z., Bennett, L. H., and Wahleithner, J. M. (2013). Fostering data literacy through preservice teacher inquiry in English language arts. *Teach. Educator* 48 (1), 8–28. doi:10.1080/08878730.2012.740151

Ball, D. L., Sleep, L., Boerst, T. A., and Bass, H. (2009). Combining the development of practice and the practice of development in teacher education. *Elem. Sch. J.* 109 (5), 458–474. doi:10.1086/596996

Ball, D. L. (1993). With an eye on the mathematical horizon: Dilemmas of teaching elementary school mathematics. *Elem. Sch. J.* 93 (4), 373–397. doi:10.1086/461730

Barnes, N., Fives, H., and Dacey, C. M. (2015). "Teachers beliefs about assessment," in *International handbook of research on teacher beliefs*. Editors H. Fives and M. G. Gill (New York: Routledge), 284–300.

Beck, J. S., Morgan, J. J., Brown, N., Whitesides, H., and Riddle, D. R. (2020). Asking, learning, seeking out: An exploration of data literacy for teaching. *Educ. Forum* 84 (2), 150–165. doi:10.1080/00131725.2020.1674438

Bell, B., and Cowie, B. (2001). *Formative assessment in science education*. Dordrecht: Kluwer.

Bernhardt, V. (2018). *Data analysis for continuous school improvement*. New York: Routledge..

Bertrand, M., and Marsh, J. A. (2015). Teachers' sensemaking of data and implications for equity. *Am. Educ. Res. J.* 52 (5), 861–893. doi:10.3102/0002831215599251

Black, P., Harrison, C., Lee, C., Marshall, B., and Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Buckingham: Open University.

Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qual. Res. Psychol.* 3 (2), 77–101. doi:10.1191/1478088706qp063oa

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educ. Meas. Issues Pract.* 30 (1), 3–12. doi:10.1111/j.1745-3992.2010.00195.x

Brown, G., Gebril, A., and Michaelides, M. (2019). Teachers' conceptions of assessment: A global phenomenon or a global localism. *Front. Edu.* 4 (16). doi:10.3389/feduc.2019.00016

Brown, G. T. L. (2008). *Conceptions of assessment: understanding what assessment means to teachers and students*. New York: Nova Science Publishers.

Brown, G. T. L., and Harris, L. R. (2009). Unintended consequences of using tests to improve learning: how improvement-Oriented resources heighten conceptions of assessment as school accountability. *J. Multidisciplinary Eval.* 6 (12), 68–91.

Bryk, A., Gomez, L., Grunow, A., and LeMahieu, P. (2015). *Learning to improve: how America's schools can get better at getting better*. Cambridge: Harvard University Press.

Caldwell, A., and Hawe, E. (2016). How teachers of years 4–8 students analyse, interpret and use information from the progressive achievement test: mathematics. *Assess. Matters* 10, 100–125. doi:10.18296/am.0019

Claudet, J. G. (2020). Using design research thinking and data-teaming processes to transform educators' professional practice: A School Improvement Case Study. *Int. J. Edu. Soc. Sci.* 7 (1), 17–41.

Cobb, P., Wood, T., Yackel, E., and McNeal, B. (1992). Characteristics of classroom mathematics traditions: An interactional analysis. *Am. Educ. Res. J.* 29, 573–604. doi:10.3102/00028312029003573

Cowie, B., and Cooper, B. (2017). Exploring the challenge of developing student teacher data literacy. *Assess. Educ. Principles Pol. Pract.* 24 (2), 147–163. doi:10.1080/0969594x.2016.1225668

Crooks, T. (2011). Assessment for learning in the accountability era: New Zealand. *Stud. Educ. Eval.* 37 (1), 71–77. doi:10.1016/j.stueduc.2011.03.002

Dalton, J., and Anderson, D. (2016). *Learning talk: Important conversations at work*: Hands On Educational Consultancy Pty Limited.

Daly, A. (2012). Data, dyads, and dissemination; Exploring data use and social networks in educational improvement. *Teach. Coll. Rec.* 114 (11), 1–38.

Datnow, A., and Hubbard, L. (2016). Teacher capacity for and beliefs about data-driven decision making: A literature review of international research. *J. Educ. Change* 17 (1), 7–28. doi:10.1007/s10833-015-9264-2

Datnow, A., and Hubbard, L. (2015). Teachers' use of assessment data to inform instruction: Lessons from the past and prospects for the future. *Teach. Coll. Rec.* 117 (4), 1–26.

Datnow, A., Park, V., and Kennedy-Lewis, B. (2012). High school teachers' use of data to inform instruction. *J. Edu. Students Placed Risk* 17 (4), 247–265. doi:10.1080/10824669.2012.718944

Datnow, A., and Park, V. (2018). Opening or closing doors for students? Equity and data use in schools. *J. Educ. Change* 19, 131–152. doi:10.1007/s10833-018-9323-6

Datnow, A., and Park, V. (2019). *Professional collaboration with purpose: teacher learning for equitable and excellent schools*. New York, NY: Routledge.

Datnow, A. (2020). The role of teachers in educational reform: A 20-year perspective. *J. Educ. Change* 21, 431–441. doi:10.1007/s10833-020-09372-5

Deneen, C., and Boud, D. (2014). Patterns of resistance in managing assessment change. *Assess. Eval. Higher Edu.* 39 (5), 577–591. doi:10.1080/02602938.2013.859654

Deneen, C. C., and Brown, G. T. L. (2016). The impact of conceptions of assessment on assessment literacy in a teacher education program. *Cogent Edu.* 3, 1225380. doi:10.1080/2331186x.2016.1225380

Education Review Office ERO (2018). Evaluation at a glance: A decade of assessment in New Zealand Primary Schools - practice and trends. Available at: https://www.ero.govt.nz/publications/evaulation-at-a-glance-a-decade-of-assessment-in-new-zealand-primary-schools-practice-and-trends/section-three-ongoing-successes-and-challenges-when-collecting-and-using-assessment/.

Edwards, F., and Ogle, D. (2021). Supporting teachers' data informed decision-making: Data informed leadership by mathematics lead teachers in New Zealand. *Teach. Dev.* 25 (1), 18–36. doi:10.1080/13664530.2020.1837217

Farrell, C. C., and Marsh, J. A. (2016). Contributing conditions: A qualitative comparative analysis of teachers' instructional responses to data. *Teach. Teach. Edu.* 60, 398–412. doi:10.1016/j.tate.2016.07.010

Fors, V., and Pink, S. (2017). Pedagogy as possibility: health interventions as digital openness. *Soc. Sci.* 6, 59–71. doi:10.3390/socsci6020059

Fulmer, G. W., Lee, I. C. H., and Tan, K. H. K. (2015). Multi-level model of contextual factors and teachers' assessment practices: An integrative review of research. *Assess. Educ. Principles, Pol. Pract.* 22, 475–494. doi:10.1080/0969594x.2015.1017445

Gearhart, M., and Osmundson, E. (2009). Assessment portfolios as opportunities for teacher learning. *Educ. Assess.* 14 (1), 1–24. doi:10.1080/10627190902816108

Gierl, M. J., Bulut, O., Guo, Q., and Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Rev. Educ. Res.* 87 (6), 1082–1116. doi:10.3102/0034654317726529

Gipps, C., and Murphy, P. (1994). *A fair test? Assessment, achievement and equity*. Buckingham, PA: Open University Press.

Gummer, E., and Mandinach, E. (2015). Building a conceptual framework for data literacy. *Teach. Coll. Rec.* 117 (4), 1–22.

Henderson, J., and Corry, M. (2020). Data literacy training and use for educational professionals. *J. Res. Innov. Teach. Learn.* [Epub ahead of print]. doi:10.1108/JRIT-11-2019-0074

Herman, J., Osmundson, E., Dai, Y., Ringstaff, C., and Timms, M. (2015). Investigating the dynamics of formative assessment: relationships between teacher knowledge, assessment practice and learning. *Assess. Edu.* 22 (3), 344–367. doi:10.1080/0969594x.2015.1006521

Hoover, N. R., and Abrams, L. M. (2013). Teachers' instructional use of summative student assessment data. *Appl. Meas. Edu.* 26 (3), 219–231. doi:10.1080/08957347.2013.793187

Hubbard, L., Datnow, A., and Pruyn, L. (2014). Multiple initiatives, multiple challenges: The promise and pitfalls of implementing data use. *Stud. Educ. Eval.* 42, 54–62. doi:10.1016/j.stueduc.2013.10.003

Jimerson, J. B., Garry, V., Poortman, C. L., and Schildkamp, K. (2020). Implementation of a collaborative data use model in a United States context. *Stud. Edu. Evaluat.* [Epub ahead of print]. doi:10.1016/j.stueduc.2020.100866

Jimerson, J. B. (2014). Thinking about data: exploring the development of mental models for "data use" among teachers and school leaders. *Stud. Educ. Eval.* 42, 5–14. doi:10.1016/j.stueduc.2013.10.010

Kippers, W. B., Poortman, C. L., Schildkamp, K., and Visscher, A. J. (2018a). Data literacy: what do educators learn and struggle with during a data use intervention? *Stud. Educ. Eval.* 56, 21–31. doi:10.1016/j.stueduc.2017.11.001

Kippers, W. B., Wolterinck, C. H. D., Schildkamp, K., Poortman, C. L., and Visscher, A. J. (2018b). Teachers' views on the use of assessment for learning and data-based decision making in classroom practice. *Teach. Teach. Edu.* 75, 199–213. doi:10.1016/j.tate.2018.06.015

Klenowski, V., and Wyatt-Smith, C. (2013). *Assessment for education: Standards, judgement and moderation*. United States: Sage.

Lai, M. K., and McNaughton, S. (2013a). "An approach for developing effective research-practice partnerships: Lessons from a decade of partnering with schools in poor urban communities," in *Research partnerships within early years education: Relational expertise and knowledges in action*. Editors J. Duncan and L. Connor (New York: Palgrave MacMillan), 49–70.

Lai, M. K., and McNaughton, S. (2013b). "Analysis and discussion of classroom and achievement data to raise student achievement," in *Data-based decision making in education: Challenges and opportunities*. Editors S. Schildkamp, M. K. Lai, and L. Earl (Netherlands: Springer), 23–48.

Lai, M., and Schildkamp, K. (2013). "Data-based decision making: An overview," in *Data-based decision making in education: Challenges and opportunities*. Editors K. Schildkamp, M. K. Lai, and L. Earl (Dordrecht, Netherlands: Springer), 9–21.

Love, N., Stiles, K. E., Mundry, S., and DiRanna, K. (2008). *A data coach's guide to improving learning for all students: Unleashing the power of collaborative inquiry*. United States: Corwin.

Mandinach, E. B., and Gummer, E. S. (2013). A systemic view of implementing data literacy in educator preparation. *Educ. Res.* 42 (1), 30–37. doi:10.3102/0013189x12459803

Mandinach, E. B., and Gummer, E. S. (2016). What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teach. Teach. Edu.* 60, 366–376. doi:10.1016/j.tate.2016.07.011

Mandinach, E. B., and Jimerson, J. B. (2016). Teachers learning how to use data: a synthesis of the issues and what is known. *Teach. Teach. Edu.* 60, 452–457. doi:10.1016/j.tate.2016.07.009

Mandinach, E. B., and Schildkamp, K. (2020). Misconceptions about data-based decision making in education: An exploration of the literature. *Stud. Edu. Evaluat.* [Epub ahead of print]. doi:10.1016/j.stueduc.2020.100842

Marsh, J. A. (2012). Interventions promoting educators' use of data: research insights and gaps. *Teach. Coll. Rec.* 114, 30–48.

Means, B., Chen, E., DeBarger, A., and Padilla, C. (2011). Teachers' ability to use data to inform instruction: challenges and supports. Report prepared for the U.S. Department of Education. Available at: www2.ed.gov/rschstat/eval/data-to-inform-instruction/report.pdf.

Means, B., Padilla, C., and Gallagher, L. (2010). Use of education data at the local level: from accountability to instructional improvement. Report prepared for the U.S. Department of Education. Available at: https://www2.ed.gov/rschstat/eval/tech/use-of-education-data/use-of-education-data.pdf.

Ministry of Education (2019). Curriculum, Progress and Achievement. Available at: https://www.education.govt.nz/our-work/information-releases/issue-specific-releases/curriculum-progress-and-achievement-programme/.

Ministry of Education (2011). *Ministry of Education position paper: Assessment [schooling sector]*. Wellington, New Zealand: Learning Media.

Ministry of Education (1993). *The New Zealand curriculum framework*. Wellington: Learning Media..

Ministry of Education (2007). *The New Zealand Curriculum mathematics*. Wellington, New Zealand: Learning Media..

Nelson, T., and Slavit, D. (2008). Supported teacher collaborative inquiry. *Teach. Edu. Q.* 35 (1), 99–116.

New Zealand Council for Educational Research ZCER. Progressive achievement Tests (PATs). Available at: https://www-nzcer-org-nz.ezproxy.waikato.ac.nz/tests/pats.

Nuthall, G. (2007). *Hidden lives of learners*. Wellington, New Zealand: NZCER Press.

Park, V., and Datnow, A. (2008). Collaborative assistance in a highly prescribed school reform model: the case of success for all. *Peabody J. Edu.* 83 (3), 400–422. doi:10.1080/01619560802222376

Penuel, W. R., Fishman, B. J., Haugan Cheng, B., and Sabelli, N. (2011). Organizing research and development at the intersection of learning, implementation, and design. *Educ. Res.* 40 (7), 331–337. doi:10.3102/0013189X11421826

Peter, M., Cowie, B., Edwards, F., Eyers, G., and Adam, A. (2017). Beyond pretty charts: From student data to pedagogical action. Available at: http://bit.ly/2q3he1R.

Pierce, R., and Chick, H. (2011). Teachers' intentions to use national literacy and numeracy assessment data: A pilot study. *Aust. Educ. Res.* 38, 433–447. doi:10.1007/s13384-011-0040-x

Raffe, C., Alonzo, D., and Loughland, T. (2019). How teachers engage with student assessment data: Understanding antecedents to data-driven decision making. *ACER Res. Conf.* [Epub ahead of print]. doi:10.13140/RG.2.2.23741.41446

Reeves, T. D., and Honig, S. L. (2015). A classroom data literacy intervention for pre-service teachers. *Teach. Teach. Edu.* 50, 90–101. doi:10.1016/j.tate.2015.05.007

Ross, E. (2017). *State teacher policy yearbook: national summary*. Washington, DC: National Council on Teacher Quality.

Ruiz-Primo, M. A., and Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *J. Res. Sci. Teach.* 44 (1), 57–84. doi:10.1002/tea.20163

Ryan, J., and Williams, J. (2007). *Children's mathematics 4–15: Learning from errors and misconceptions*. Maidenhead, UK: Open University Press.

Schildkamp, K., and Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teach. Teach. Edu.* 26 (3), 482–496. doi:10.1016/j.tate.2009.06.007

Schildkamp, K., and Poortman, C. L. (2015). Factors influencing the functioning of data teams. *Teach. Coll. Rec.* 117 (4), 1–42.

Schildkamp, K., Poortman, C., Luyten, H., and Ebbeler, J. (2017). Factors promoting and hindering data-based decision making in schools. *Sch. Effectiveness Sch. Improvement* 28 (2), 242–258. doi:10.1080/09243453.2016.1256901

Shepard, L. A. (2019). Classroom assessment to support teaching and learning. *ANNALS Am. Acad. Polit. Soc. Sci.* 683 (1), 183–200. doi:10.1177/0002716219843818

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harv. Educ. Rev.* 57 (1), 1–22. doi:10.17763/haer.57.1.j463w79r56455411

Spillane, J. P. (2012). Data in practice: conceptualizing the data-based decision-making phenomena. *Am. J. Edu.* 118 (2), 113–141. doi:10.1086/663283

Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. Abingdon: Routledge.

US Department of Education (2015). Every student Succeeds act (ESSA). Available at: https://www.ed.gov/essa.

Van Gasse, R., Vanlommel, K., Vanhoof, J., and Van Petegem, P. (2020). Teacher interactions in taking action upon pupil learning outcome data: A matter of attitude and self-efficacy? *Teach. Teach. Edu.* 89 102989, doi:10.1016/j.tate.2019.102989

Van Gasse, R., Vanlommel, K., Vanhoof, J., and Van Petegem, P. (2017). The impact of collaboration on teachers' individual data use. *School Effectiveness and School Improvement. Int. J. Res. Pol. Pract.* 28 (3), 489–504. doi:10.1080/09243453.2017.1321555

Vanlommel, K., Van Gasse, R., Vanhoof, J., and Van Petegem, P. (2017). Teachers' decision-making: data based or intuition driven? *Int. J. Educ. Res.* 83, 75–83. doi:10.1016/j.ijer.2017.02.013

Visscher, A. J. (2020). On the value of data-based decision making in education: The evidence from six intervention studies. *Stud. Edu. Evaluat.* [Epub ahead of print]. doi:10.1016/j.stueduc.2020.100899

Volante, L., DeLuca, C., Adie, L., Baker, E., Harju-Luukkainen, H., Heritage, M., et al. (2020). Synergy and tension between large-scale and classroom

assessment: International trends. *Edu. Measure. Iss. Prac.* 39 (4), 21–29. 10.1111/emip.12382

Wayman, J. C., and Jimerson, J. B. (2014). Teacher needs for data-related professional learning. *Stud. Educ. Eval.* 42, 25–34. doi:10.1016/j.stueduc.2013.11.001

Wiener, R., and Hall, D. (2004). Accountability under no child left behind. *Clear. House* 78 (1), 17–21. doi:10.3200/TCHS.78.1.17-21

Windschitl, M., Thompson, J., Braaten, M., and Stroupe, D. (2019). Sharing a vision, sharing practices: how communities of educators improve teaching. *Remedial Spec. Edu.* 40 (6), 380–390. doi:10.1177/0741932518810796

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

frontiers in Education

ORIGINAL RESEARCH
published: 07 May 2021
doi: 10.3389/feduc.2021.616216

Changes in Student Motivation and Teacher Decision Making When Implementing a Formative Assessment Practice

Gunilla Nässtrom[1], Catarina Andersson[2,3], Carina Granberg[3,4], Torulf Palm[2,3] and Björn Palmberg[2,3]*

[1]Department of Education, Umeå University, Umeå, Sweden, [2]Department of Science and Mathematics Education, Umeå University, Umeå, Sweden, [3]Member of Umeå Mathematics Education Research Centre (UMERC), Umeå University, Umeå, Sweden, [4]Department of Applied Educational Science, Umeå University, Umeå, Sweden

Motivation is a prerequisite for students' learning, and formative assessment has been suggested as a possible way of supporting students' motivation. However, there is a lack of empirical evidence corroborating the hypothesis of large effects from formative assessment interventions on students' autonomous forms of motivation and motivation in terms of behavioral engagement in learning activities. In addition, formative assessment practices that do have an impact on students' motivation may put additional requirements on teachers than more traditional teaching practices. Such requirements include decisions teachers need to make in classroom practice. The requirements on teachers' decision-making in formative assessment practices that have a positive impact on students' autonomous forms of motivation and behavioral engagement have not been investigated. This study describes one teacher's formative assessment practice during a sociology course in upper secondary school, and it identifies the requirements for the teacher's decision-making. The teacher had participated in a professional development program about formative assessment just prior to this study. This study also investigated changes in the students' motivation when the teacher implemented the formative assessment practice. The teacher's practice was examined through observations, weekly teacher logs, the teacher's teaching descriptions, and an interview with the teacher. Data on changes in the students' type of motivation and engagement were collected in the teacher's class and in five comparison classes through a questionnaire administered in the beginning and the end of the course. The students responded to the questionnaire items by choosing the extent to which they agreed with the statements on a scale from 1–7. The teacher's formative assessment practice focused on collecting information about the students' knowledge and skills and then using this information to make decisions about subsequent instruction. Several types of decisions, and the knowledge and skills required to make them that exceed those required in more traditional teaching practices, were identified. The students' in the intervention teacher's class increased their controlled and autonomous forms of motivation as well as their engagement in learning activities more than the students in the comparison classes.

Keywords: formative assessment, student motivation, teacher decision-making, student engagement, classroom practice

OPEN ACCESS

Edited by:
Chris Davison,
University of New South Wales, Australia

Reviewed by:
Susan M Brookhart,
Duquesne University, United States
Anil Kanjee,
Tshwane University of Technology, South Africa

****Correspondence:***
Gunilla Nässtrom
gunilla.nasstrom@umu.se

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

Received: *11 October 2020*
Accepted: *23 April 2021*
Published: *07 May 2021*

Citation:
Nässtrom G, Andersson C, Granberg C, Palm T and Palmberg B (2021) Changes in Student Motivation and Teacher Decision Making When Implementing a Formative Assessment Practice. Front. Educ. 6:616216. doi: 10.3389/feduc.2021.616216

INTRODUCTION

Motivation

Motivation is the driving force of human behavior and is a prerequisite for students' learning. Students' motivation to learn may be manifested through students' *behavioral engagement*, which refers to how involved the student is in learning activities in terms of on-task attention and effort (Skinner et al., 2009). Research has consistently found student reports of higher levels of behavioral engagement to be associated with higher levels of achievement and less likelihood to drop out of school (Fredricks et al., 2004).

Students may also have different types of motivation. That is, they may be motivated for different reasons (Ryan and Deci, 2000). Students who have *autonomous forms of motivation* engage in learning activities either because they find them inherently interesting or fun, and feel competent and autonomous during the activities, or because they find it personally valuable to engage in the activities as a means to achieving positive outcomes. Students with *controlled forms of motivation*, on the other hand, experience the reasons for engaging as imposed on them. They may feel pressured to engage because of external rewards (such as being assigned stars or monetary rewards), to avoid discomfort or punishment (such as the teacher being angry or assigning extra homework), to avoid feeling guilty (e.g., to avoid the feeling of letting parents or the teacher down), or to attain ego enhancement or pride. Students' type of motivation has consequences for their learning. Autonomous forms of motivation have been shown to be associated with greater engagement, but also with higher-quality learning and greater psychological well-being. The more-controlled forms of extrinsic motivation, on the other hand, have been shown to be associated with negative emotions and poorer coping with failures (Ryan and Deci, 2000). Successfully supporting student motivation is, however, not an easy task, and several studies have shown that student motivation often both decreases and becomes less autonomous throughout the school years (Winberg et al., 2019).

Formative Assessment as a Means of Supporting Students' Motivation

Formative assessment, which is a classroom practice that identifies students' learning needs through assessments and then adapts the teaching and learning to these needs, has been suggested as a possible way of supporting student motivation (e.g., Clark, 2012). There is great variation in how scholars conceptualize formative assessment, and Stobart and Hopfenbeck (2014) describe some common conceptualizations. Formative assessment practices may be teacher-centered, student-centered, or a combination of these. Teacher-centered approaches to formative assessment focus on the teachers' actions, and in these practices teachers gather evidence of student learning, for example, through classroom dialogue or short written tests, and they adapt feedback or the subsequent learning activities to the information gathered from these assessments. In the student-centered approaches to formative assessment, the students are involved in peer assessment and peer feedback and/or self-assessment in order to take a more proactive role in the core formative assessment processes of identifying their learning needs and acting on this information to improve their learning. It may be noted however, that although the teacher may be seen as the proactive agent in teacher-centered formative assessment practices, such practices may still have the students at the center in the sense that the focus is on identifying the students' learning needs and adapting the classroom practices to these needs. The following definition by Black and Wiliam (2009) incorporates many of the meanings given to formative assessment:

> Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited (p. 9).

Researchers have provided different suggestions about how formative assessment may affect student motivation. Brookhart (2013) emphasizes the students as the main source and users of learning information in formative assessment, and points to the consistency between the formative assessment cycle [establishing where the learners are in their learning, establishing where they are going (the learning goals), and establishing what needs to be done to get there] (Wiliam and Thompson, 2008), and the phases of self-regulated learning (Zimmerman, 2000). Brookhart (2013) uses several motivation theories to discuss how the characteristics of assessment tasks, the classroom environments they are administered in, and teachers' feedback may influence students' motivation to use assessment information and engage in learning activities. Heritage and Wylie (2018) emphasize the inclusion of both teachers and students as active participants in the formative assessment processes. In these processes teachers and students notice the students' learning, respond to this information by choosing learning tasks suitable for the students to take the next steps in their learning, and provide feedback that emphasize evidence of students' learning. They argue that such processes support the development of an identity as effective and capable learners. Such identity beliefs enhance students' motivation to engage in learning activities. Shepard et al. (2018) argues that formative assessment practices in which feedback helps students to see what they have learned and how to improve may foster a learning orientation. Within such a learning orientation, students find it personally valuable to engage in learning activities, and they feel less controlled in their motivation by, for example, a need to please others or appear competent. Pat-El et al. (2012) argue that teacher feedback that helps students monitor their learning progress and provides support for how goals and criteria can be met, may enhance students' satisfaction of the three psychological needs of competence, autonomy and relatedness. Pat-El et al. (2012) then draw on self-determination theory (Ryan and Deci, 2000), which posits that these psychological needs

influence students' autonomous forms of motivation. Their findings in a questionnaire study performed on one single occasion (Pat-El et al., 2012) indicated that competence and relatedness mediated an influence of feedback on autonomous motivation. In an intervention study by Hondrich et al. (2018), teachers implemented teacher-centered formative assessment practices involving the use of short written tasks to assess students' conceptual understanding, feedback and adaptation of instruction. They found an indirect effect of formative assessment on autonomous motivation mediated by perceived competence (the other two psychological needs, competence and relatedness, were not included in the study). Thus, when discussing the possible mechanisms by which formative assessment may affect students' engagement and type of motivation, some scholars focus on the students as the main proactive agents in the formative assessment processes (e.g., Brookhart, 2013), while others also focus on the teachers (e.g., Heritage and Wylie, 2018). Researchers also use different theories of motivation when discussing the nature of these possible mechanisms. Shepard et al. (2018) uses learning orientation theory, while Pat-El et al. (2012) and Hondrich et al. (2018) draw on self-determination theory when discussing possible mechanisms for the effects of formative assessment on students' type of motivation. When discussing the effects of formative assessment on students' engagement, Heritage and Wylie (2018) use the notion of identity beliefs, while Brookhart (2013) uses several different motivation theories.

However, as of yet there is no solid research base corroborating neither the promising hypothesis of large effects, nor the mechanisms underlying such effects, from different approaches to formative assessment on student motivation in terms of autonomous motivation or behavioral engagement in learning activities. Studies that investigate effects of formative assessment on motivation within an ecologically valid, regular classroom environment are scarce (Hondrich et al., 2018). Indeed, we performed a comprehensive literature search in the databases ERIC, APA PsychInfo, Academic Search Premier and SCOPUS, which returned 557 journal articles, but only 11 of them were empirical studies that examined the association between formative assessment and grade 1–12 students' type of motivation or engagement in learning activities. In the database search we used the Boolean search command (motivation OR engagement) AND ("formative assessment" OR "assessment for learning" OR "self-assessment" OR "peer-assessment" OR "peer feedback") AND (effect* OR impact OR influenc* OR affect* OR relation OR predict*) AND (school* OR grade* OR secondary OR primary OR elementary) in the title, abstract and keywords. We included a number of terms commonly used for formative assessment. The term "motivation" was used to include all articles that deal with the motivational terms investigated in our own study, and the term "engagement" was used as a complement to the term "motivation" to ensure that articles focusing engagement were found. Since we were interested in empirical evidence for the association between formative assessment and motivational outcomes, a number of terms commonly used in such studies were used to limit the search to such studies. Finally, since the present study included upper-secondary school students, and university studies may differ in important aspects from compulsory school, we limited our search to journal articles including studies from school year 1–12. A number of terms (see above) were used to filter out studies involving higher education or out-of-school learning.

In the literature search, we found a few studies that are based on questionnaire responses from students on one single occasion, and they have found associations between students' *autonomous motivation* and different approaches to formative assessment. Pat-El et al. (2012) and Federici et al. (2016) found an association between students' autonomous motivation and the students' perceptions of teachers' providing feedback that facilitates the monitoring of their learning progress and an understanding of how goals and criteria can be met (formative feedback). However, in the context of portfolio use, Baas et al. (2019) found an association between autonomous motivation and students' perceptions that scaffolding is integrated in their classroom practice, but not between autonomous motivation and the monitoring of growth. Gan et al. (2019) found associations between students' autonomous motivation and their perceptions of a daily classroom practice involving continuous informal assessments and dialogic feedback, and Zhang (2017) found associations between autonomous motivation and students' perceptions of their possibilities to self-assess and take follow-up measures in the classroom practice. Thus, the three first studies focused on teacher-centered formative assessment. The studies by Pat-El et al. (2012) and Federici et al. (2016) both involved an emphasis on the teachers' feedback, while the study by Baas et al. (2019) did so in the specific context of portfolio use. The fourth study (Zhang, 2017) focused on student-centered formative assessment practices in terms of self-assessment.

Intervention studies examining the effects on students' type of motivation show mixed effects. Three intervention studies (1–6 months) were primarily teacher-centered and focused on both the collection of evidence of student learning and feedback. In these interventions, tests or assignments and educational materials were made available for the teachers or students to use. When teachers were not provided with information about how to best use information about students' progress for learning purposes, the intervention did not have an effect on students' autonomous motivation (Förster and Souvignier, 2014). When the teachers were provided with a short professional development course (13 h; Hondrich et al., 2018) or a digital formative assessment tool (Faber et al., 2017) to aid the formative processes of providing student assignments and feedback, then small effects were found on students' self-reported autonomous motivation. Finally, in the study by Meusen-Beekman et al. (2016) the teachers were provided with information about a larger array of approaches to formative assessment. This information included how to establish and share assessment criteria with students, how to implement rich questioning and provide feedback, and how to support either peer assessment or self-assessment (two different intervention conditions). Both of these intervention conditions showed a nearly medium-size effect on autonomous motivation in comparison to a control group,

and the teachers' practices in the two conditions can be characterized as both teacher-centered and student-centered.

Three studies have investigated the association between formative assessment and students' motivation in terms of *engagement in learning activities*. One study focused on teacher feedback (Federici et al. (2016), which can be considered a teacher-centered approach to formative assessment. The second study (Wong, 2017) involved a student-centered approach, and the formative assessment practice in the third study (Ghaffar et al., 2020) can be characterized as both teacher-centered and student-centered. Federici et al. (2016) analyzed questionnaire responses from students on one single occasion, and found an association between students´ perceptions of teachers' formative feedback and students' persistence when doing schoolwork. Wong (2017) found a medium-size effect on students' self-reported engagement from an intervention in which the researcher taught self-assessment strategies to the students (thus no teacher professional development was necessary). In the study by Ghaffar et al. (2020) a teacher engaged her students in co-construction of writing rubrics together with both teacher feedback and peer feedback. The results indicated some positive outcomes for students' autonomous motivation and engagement in learning activities in comparison with a control class during the two-months intervention using a writing assignment.

Thus, formative assessment may be carried out in a range of different ways, and the few existing studies investigating its effects on motivation have been built on formative assessment practices with different characteristics. Because formative assessment practices may have different characteristics, which in turn may affect students' motivation differently, the available research base needs to be extended with more investigations into the effects of all of the different main approaches to formative assessment (such as student self-assessment or more teacher-centered approaches where the teachers carry out the assessment and provide feedback) to be able to draw more well-founded conclusions regarding the effects of formative assessment on motivation. In particular, there exist only a very few intervention studies investigating the effects of teacher-centered formative assessment practices on autonomous motivation, and we have not found any intervention studies investigating the effects of teacher-centered formative assessment practices on students' behavioral engagement.

Requirements on Teachers' Decision-Making in Formative Assessment Practices

Studying the effects of different classroom practices on student outcomes such as motivation is important, for obvious reasons, but studying what is required by teachers to carry out these practices is also of significant value. Teachers need to master many different skills in order to carry out their teaching practices, and different practices may require a slightly different set of skills. However, regardless of the type of practice, the importance of teachers' decision-making while planning and giving lessons has been recognized for a long time (for reviews, see for example Shavelson and Stern, 1981; Borko et al., 2008; Hamilton et al., 2009; Datnow and Hubbard, 2015; Mandinach and Schildkamp, 2020), and teachers' decision-making may be regarded to be at "the heart of the teaching process" (Bishop, 1976, p. 42). Teachers' decision-making is seldom straightforward, however. Teachers need to make judgments and decisions in a complex, uncertain environment, having limited time to process information (Borko et al., 2008) and, in general, having limited access to information. Teachers' decisions about content, learning activities, and so forth are affected by a number of variables such as their knowledge, beliefs, and goals (Schoenfeld, 1998) that are shaped by the context in which they reside. The types of decisions teachers need to make, that is, the requirements on the teachers' decision making, depends on the type of classroom practice that is carried out. For example, if a classroom practice includes adapting teaching to students' learning needs, then decisions about gathering information about these needs and how to adapt teaching to these needs have to be made. If a classroom practice does not have such an adaptation focus, then the teacher may not be required to make these kinds of decisions to the same extent. Furthermore, the knowledge and skills the teachers need to have to make decisions depends on the type of decisions they have to make. In the classroom practice with the adaptation focus, teachers need the knowledge and skills to successfully use assessment information to make decisions on teaching adaptations that fit different student learning needs.

Teachers' teaching during lesson has been characterized as carrying out well-established routines (Shavelson and Stern, 1981). The routines include monitoring the classroom, and if the routine is judged to be proceeding as planned there is no need to deviate from the lesson plan. But, if the teacher sees cues that the lesson is deviating too much from the plan, then the teacher has to decide whether other actions need to be taken. The main issue for many teachers in their monitoring seems to be the flow of the activity, that is, the decisions are most often based on the students' behavior such as their lack of involvement or other behavioral student problems (Shavelson and Stern, 1981), and teachers seldom use continuous assessments of students' learning as a source of information when deciding how to resolve pedagogical issues (Lloyd, 2019).

Practices such as formative assessment, in which teachers make decisions based on assessment of students' subject matter knowledge, may require other types of teacher knowledge and skills than in practices in which decisions are primarily based on teachers' needs to cover the curriculum, their experiences with former students, current students' prior learning, their intuition, and the behavior in the classroom. It has been argued that in formative assessment practices, teachers need to be skillful in a variety of ways in order to gather information about students' subject matter knowledge, how to interpret the students' responses in terms of learning needs, and how to use these interpretations to adapt the classroom practice to improve the students' learning (Brookhart, 2011; Means et al., 2011; Gummer and Mandinach, 2015; Datnow and Hubbard, 2016). Consequently, there have been calls for developing support

for teachers on not only how to gather information about students' learning, but also on how to interpret the collected information and how to use these interpretations for instructional purposes (Mandinach and Gummer, 2016). However, despite many attempts at professional developments aimed at building teachers' capacity for using assessment data when making instructional decisions, many teachers often feel unprepared to do so (Datnow and Hubbard, 2016), and many professional development programs in formative assessment have failed to lead to substantially developed formative assessment practices (Schneider and Randel, 2010). If teachers are supposed to implement new practices, they need the knowledge and skills required to do so. If they do not already possess them they need to be provided with sufficient support to acquire them. Teachers will not implement new practices they do not find viable to carry out, but to be able to provide teachers with necessary support, and in order for teachers to be able to assess the viability of implementing the new practice, insights into the skills necessary to carry out the practice are needed.

Therefore, studies that describe the decisions teachers are required to make, and the skills needed to make these decisions, in order for classroom practices to have a positive effect on students' motivation would be of fundamental value. Some studies have explored the decisions teachers make in practices that include aspects of formative assessment. For example, Hoover and Abrams (2013) explored teachers reported use of summative assessments in formative ways. They found that most teachers reported use of summative assessment data in order to change the pace of instruction, to regroup or remediate students as needed, or to provide instruction using different strategies. However, a minority of the teachers made such decisions on a weekly basis, and the decisions were most often based on central tendency data, interpretation of results within the context of their teaching or validation of test items. Such instructional decisions would be informed by conclusions about students' areas of weaknesses, but less on conclusions about students' conceptual understanding (Oláh et al., 2010). However, in the literature search described in *Formative Assessment as a Means of Supporting Students' Motivation* section, we found no studies that describe teachers' decision-making in daily formative assessment classroom practices that are empirically shown to have positive effects on students' autonomous forms of motivation or behavioral engagement.

RESEARCH QUESTIONS

In the present study we analyze the characteristics of a teacher's implemented teacher-centered formative assessment practice, including the practice's requirements on the teacher's decision-making. We also investigate the changes in the students' motivation, both in terms of engagement in learning activities and in terms of the type of motivation. We ask the following RQs:

RQ1. What are the characteristics of this teacher-centered formative assessment practice, and what are the requirements on the teacher's decision-making?

RQ2. Does the intervention class students' behavioral engagement in learning activities increase in comparison with five comparison classes?

RQ3. Does the intervention class students' type of motivation (autonomous and controlled forms of motivation) increase in comparison with the five comparison classes?

METHODS AND MATERIALS

Participants

The intervention teacher, Anna (fictitious name), with comprehensive university studies in all her teaching subjects and extensive teaching experience (>20 years), had participated in a professional development program in formative assessment the year before this intervention. During that year she worked in another school than the school in which the intervention took place, and the principal at that school had decided that all teachers in the school had to attend the professional development program. Thus, she had not volunteered to participate in the program, nor had she been selected to the program based on any of her characteristics. At the beginning of the autumn term in 2016, she started teaching a course in sociology with 19 second-year students who were enrolled in the Child and Recreation Program at a Swedish upper secondary school. Inspired by the professional development program, Anna implemented a formative assessment practice in her class during October 2016 to May 2017. All 19 students in the intervention class were invited to take part in the study, and none of the students declined to participate. Twelve students attended class on both occasions when the student questionnaires were administered.

The students in the comparison classes were all taught by experienced teachers. These teachers had not participated in the professional development program in formative assessment and did not specifically aim to implement a formative assessment practice. The students in these classes were enrolled in the Building and Construction Program, the Industrial Technology Program, the Child and Recreation Program, and the Social Science Program (two classes). All programs are vocational programs except for the Social Science Program, which is an academic program. The comparison classes were chosen based on the fact that the classes in these programs, despite program differences, did not differ much in overall academic achievement when they began their upper-secondary school studies (students enrolled in the Social Science Program had a little higher grade-average from school-year 9, which is the school year that precedes upper secondary school, than the students from the classes in the other three programs). In Sweden a number of courses are taken at the same time, but during the period when the intervention class took the sociology course none of the other classes took the exact same course. Therefore, type of motivation and behavioral engagement of the students in the comparison classes were measured in the courses most similar to the sociology course. These courses were a social science course for the two classes belonging to the Social Science Program, and a history course for the classes belonging to The Industrial

Technology program, the Building and Construction Program, and the other class in the Child and Recreation Program. All of these courses belong to the social science domain, and both the social science course and the sociology course include an historical perspective. The courses (including the sociology course taken by the intervention class) corresponded to five weeks of full-time studies, but since the students take several courses at the same time they lasted through the whole intervention period. Although program-specific courses differ between programs, the same academic course, for example in social science, is not dependent on the program. Among the 121 students in the comparison classes, 72 of them agreed to participate as well as attended class on both occasions the questionnaires were administered so they could complete them. Only three of the 121 students declined to participate. The participating students, in total 84, were 17–18 years of age and enrolled in the same upper secondary school. Among the students in the intervention group, 55% were girls and 86% had Swedish as their mother tongue. In the comparison group, 50% were girls and 88% had Swedish as their mother tongue.

The research project was conducted in accordance with Swedish laws as well the guidelines and ethics codes from the Swedish Research Council that regulate and place ethical demands in the research process (http://www.codex.vr.se/en/). For the type of research conducted in this study, it is not necessary to apply for ethical evaluation to the Swedish Ethical Review Authority. Written consent was obtained from the teacher and the students.

Data Collection and Method of Analysis for RQ1

For RQ1, multi-method triangulation was used with four qualitative methods: classroom observation, the teacher's logs, the teacher's teaching descriptions and a teacher interview. The aim of this triangulation was to develop a comprehensive view of the teacher's formative assessment practice and the decisions she made when carrying out this practice. This type of multi-method triangulation is a way of enhancing internal validity of the qualitative data (Meijer et al., 2002). The intention was not to establish if the data from these methods would show the same results. Classroom observations can provide examples of how the teacher uses formative assessment in the classroom, but a single observation cannot show the variation of the practice over different lessons or how common an observed practice is. Teacher logs, teacher teaching descriptions and teacher interviews, on the other hand, may provide more information about how a classroom practice varies over lessons and how common certain aspects of the practice are. Therefore, the four methods were used to provide complementary data on the teacher's formative assessment practice.

Teacher Logs and Teacher's Teaching Descriptions

The teacher log was used over a period of 6 months, from November 2016 to April 2017. Anna was asked to make notes shortly after each lesson or series of lessons in the intervention class. The log was digital and asked, with six questions, Anna for information about each teaching activity used during the lesson. The questions asked for a description of the implementation of the activity, information about whether the activity involved any pedagogical adjustments, the rationale behind the decision to choose each activity, an evaluation of the implementation, an evaluation of the outcome of the activity, and finally the log included an open question for further comments. During that period Anna wrote detailed notes answering the six questions in the teacher log twelve times. Some of these described lessons entailed more than one teaching activity so data consisted of written reports from 17 teaching activities.

Anna furthermore wrote teaching descriptions (5–10 pages) at three occasions; before, in the middle, and at the end of the intervention. These descriptions aimed to capture her overall teaching design, how her teaching with respect to formative assessment changed over time, and her rationale for her teaching decisions.

Classroom Observation

One classroom observation was conducted in February 2018, just prior to the teacher interview, by the first author. The researcher used a protocol to keep notes of Anna's teaching, aiming for observing what activities Anna implemented, how she introduced them, if the students reacted as if they were used to the activity or not, and how the students engaged in that particular activity. The researcher furthermore informally spoke to the teacher before the observation, and information gained from this conversation was also included in the field notes.

Teacher Interview

The interview was conducted by the first author immediately after the classroom observation. It lasted about 1 h, and was audio recorded. To begin with, Anna was asked to describe her teaching before the intervention, especially activities that she had changed or excluded when she planned for the intervention. She was thereafter invited to, in detail, describe each of her chosen activities used during the intervention. Activities written down in the teacher log and noticed during the observation were also brought up during the interview to be described in more detail. She was then asked to describe her motives behind the decisions to change, exclude, or choose a particular activity. She was finally invited to elaborate on how she thought the activity could work as a part of a formative classroom practice and how she expected the activity to support her students' learning.

Method of Analysis

To capture the characteristics of Anna's formative assessment practice and the requirements from this practice on Anna's decision making, the analysis was conducted in three steps. The analysis was made jointly by the first, third and fourth author, and decisions on categorizations were made in consensus.

The first step aimed to capture Anna's classroom practice before and during the formative assessment intervention. This was done by analyzing the field notes from the classroom observation, the logbooks, the teacher's teaching descriptions and the interview data to identify learning activities that were regularly implemented. Thereafter, the definition by Black and

TABLE 1 | Example of the analysis.

Data – Activity	Analysis
Excerpt from the interview: "Now I start my teaching of each topic by using Google forms. My students are asked to explain important concepts they need to learn. This gives me a pretty good picture of their prior knowledge and I can adjust my teaching accordingly. …. I need this information to choose where to start, what to focus on and what pace to choose. When they have given their answers in Google forms we can also look at the classes' answers and the diagrams of what they have answered. We can then discuss the concepts directly, and they learn much when we discuss the concepts at the same time as they see what they have answered." *Excerpt from the teacher log:* Today I used my favorite tool Google forms to start the sociology course. I had chosen 17 concepts concerning Marx and Durkheim and the students could choose from four answers: "Never heard of." "Have heard but can't explain," "Can explain to some extent," "I know for sure." I think this is a good way of gathering information about students' prior knowledge. It helps me plan the forthcoming weeks. But it is also good to take the opportunity to choose tricky concepts to discuss directly after they have submitted their answers. *Excerpt from the teachers' teaching description:* I always use Google forms to ask students questions before we start a topic, for example "Politics." The students' answers will give me information about their prior knowledge. Information that I then use to plan my teaching. I chose concepts I think are essential to understanding the course, and create questions that will make them describe their understanding. [A table is inserted in her description with the sentences the students were asked to complete: Government decides . . . , Parliament decides . . . , and questions like: Who works in the Government? Who works in the Parliament? . . . and so forth]. After they have submitted their answers we discuss them together, and I ask them questions, making them clarify. This is also a good learning opportunity.	*Formative classroom practice*: The activity is being carried out at the beginning of each topic or a course and is therefore interpreted as regularly carried out and part of her regular classroom practice. The activity is categorized as formative assessment since Anna elicits information about the students' learning needs and modifies her teaching accordingly. *Decision-making* Before the lesson introducing a topic or a course Anna needs to make a couple of decisions. She needs to decide what kind of information she needs to elicit from her students to gain insights of their prior knowledge. For example, their understanding of key concepts in politics, sociology, and so forth. Thereafter she needs to decide what questions to ask that will provide that information. For example, to assess their own understanding of concepts in sociology, to complete sentences or answer questions concerning politics. She finally needs to decide how the questions and responses should be administered – in this case, using Google forms. When her students answer these questions during the lesson there are a couple of decisions she needs to make instantaneously. How to interpret the responses to identify students' learning needs and how to act, for example, to provide feedback accordingly. The students' answers will furthermore provide information that Anna intends to use to plan forthcoming lessons. That is, to make long-term decisions on how to adjust future teaching according to students' prior knowledge and learning needs. *Teacher's knowledge and skills:* When Anna makes her decision about what information about students' prior knowledge she wants to ask for, she will need comprehensive knowledge about the subject, in this case sociology and politics. She furthermore needs knowledge on how to choose and construct questions to gain that kind of information. Her decision to use Google forms will put demands on her technical skills. Finally, Anna needs knowledge about how to interpret students' responses, how to identify learning needs and to choose actions accordingly. Both instantly during the lesson as well as for planning future lessons.

Wiliam (2009, p. 9) quoted in *Formative Assessment as a Means of Supporting Students' Motivation* section, was used as a framework for examining which of the identified activities that could be characterized as being formative assessment. Thus, activities in which the teacher or students elicited evidence of student achievement, and used this information to make decisions on the next step in the teaching or learning practice, would be categorized as formative assessment. In the final step of the analysis the collected data was used to identify the types of decisions Anna's practice required her to make, and the knowledge and skills needed to make these decisions. **Table 1** provides examples of this analysis procedure.

Data Collection and Method of Analysis for RQ2 and RQ3

For RQ2 and RQ3, a quasi-experimental design with intervention and comparison classes was used. The participating students completed a web questionnaire at the beginning and the end of the intervention. They did so during lesson time and on each occasion one of the authors was there to introduce the questionnaire and answer questions from the students if anything was unclear to them. This method of data collection will be further described in the following.

Questionnaires

Measures of changes in student engagement and type of motivation in both the intervention class and comparison classes were obtained through a questionnaire administered before and after the intervention. All items measuring students' engagement in learning activities were statements that the students were asked to mark to what extent they agreed with on a scale from 1 (not at all) to 7 (fully agree). The items measuring students' type of motivation were statements of reasons for working during lessons or for learning the course content. The students were asked to mark to what extent these reasons were important on a scale from 1 (not at all a reason) to 7 (really important reason). Five items measuring behavioral engagement were adaptations of items from Skinner et al. (2009) questionnaire items on behavioral engagement, and six items each measuring autonomous and controlled motivation were adapted from Ryan and Connell (1989) Self-Regulation Questionnaire. The adaptations were made to suit the context of the participants, and before the study these adaptations were piloted with students in four other classes of the same age group to ensure that the questions were easy to understand. A list of all questionnaire items can be found in Appendix A. An example of a behavioral engagement item is: "I am always focused on what I'm supposed

to do during lessons." Examples of items measuring autonomous and controlled motivation are: "When I work during lessons with the tasks I have been assigned, I do it because I want to learn new things" and "When I try to learn the content of this course, I do it because it's expected of me." Cronbach's alpha for each set of the items in spring/fall was 0.86/0.88 for behavioral engagement, 0.90/0.89 for autonomous motivation, and 0.74/0.78 for controlled motivation, indicating acceptable to good internal consistency of the scales. To examine whether each scale was unidimensional, exploratory factor analysis was performed on each set of items for each time point. The extraction method was principal axis factor and the scales were deemed to be unidimensional if the scree plot had a sharp elbow after the first factor, if the eigenvalue of the second factor was <1, and if parallel analysis suggested that only one factor should be retained. The choice of not doing exploratory factor analysis on all items for each time point was based on that the low subject to item ratio (<5:1) would make the risk of misclassifying items and not finding the correct factor structure high (Costello and Osborne, 2005). The mean of the items connected to a construct at each time point was used as a representation of students' behavioral engagement, autonomous motivation, and controlled motivation at the time point.

Statistical Analysis

To investigate the changes in students' behavioral engagement (RQ2) and autonomous and controlled motivation (RQ3), mean differences in the responses to the questionnaire items pertaining to these constructs between fall and spring were calculated for both students in the intervention class and in the comparison classes. Students are nested within classes, and therefore it was not reasonable to treat the comparison classes as one group. Because of this, the change in each construct between fall and spring for the intervention class was compared with the same change in each of the comparison classes. Partially due to nesting of students within classes, the study lacks power and statistically significant differences in mean values (or variances) were not seen between groups. We therefore chose to indicate the size of the difference in changes between the intervention class and the comparison classes through calculation of Hedges' g (Hedges, 1981). A commonly used interpretation of sizes of this type of effect measure suggested by Cohen (1988) is that 0.2, 0.5, and 0.8 indicate small, medium, and large effects, respectively.

RESULTS

Characteristics of the Formative Assessment Practice and the Requirements on the Teacher's Decision-Making

Teaching Practice Before the Intervention

Anna's teaching before the professional development program can be characterized as traditional, as Anna described:

> My lessons followed the same dramaturgy. I started by presenting the aim of the present lesson by writing on the smart board or presented as the first slide of my PowerPoint. Thereafter I gave a lecture for 20–30 minutes using my PowerPoint. Then, the students worked with assignments, individually or in groups. Sometimes we watched an educational film followed up by a whole class discussion. . . . I always tried to choose films, questions and tasks that I believed would be interesting for my students to work with.

Anna described her way of interacting with her students, that she, during students' work, mainly supported students who asked for help. Anna expressed the challenges of providing support to 30 students in the classroom: "It is difficult to divide my time wisely . . . the students that are active and ask for help get more support than those not reaching out for me, and these students also need help." When Anna is asked to describe her assessment practice she described her way of using written tests and reports mainly for summative purposes, grading the students.

Decisions, knowledge and skills

Anna's classroom practice entailed some recurrent decisions. For example, to decide on how to present subject matter in a way that the students would understand and find engaging. Anna based these decisions on her general knowledge of teenagers' interests. Before the intervention, Anna had decided to primarily help students who asked for support, which is a decision she questioned during the intervention since she knows that students who really need help don't always ask for it. Since Anna's assessment focus was on summative assessment, her assessment decisions pertained to these kinds of assessments, and she needed skills to assess students' gained knowledge in relation to national standards and to decide on the assignment of grades to the students. Decisions rarely concerned how to gather and interpret information about the current students' knowledge and skills in order to use this information to support their learning. Thus, she did not need skills to make such decisions. When she planned and carried out her teaching, judgments of what the students would understand were based on her knowledge about the content that had been included in prior courses the students had taken (and thus should have been learned) and experiences of former students' understanding of the content in the current course.

Teaching Practice During the Intervention – A Formative Assessment Practice

During the professional development program, Anna changed her view of teaching from a focus on how to teach for the students to be interested in learning the subject to a focus on the students' actual learning, as Anna describes:

> I used to aim for planning interesting lessons, my idea was that if students are interested they will be motivated . . . I was not particularly interested in each student's learning besides at the end of a course, when I assessed their level of knowledge . . . However, I now realize that I gave my students assignments that were too difficult

> (even though interesting). They were not familiar with the essential concepts they needed to solve the tasks.

After the professional development program, Anna said she started to ask herself three questions: 1) What are the students' knowledge and skills in relation to the learning goals at the beginning of a teaching and learning unit? 2) What are their knowledge and skills later on during learning sequences? 3) Based on the answers to 1) and 2), what would be the best teaching method to meet these learning needs? This change in view had important consequences for her practice, including her decision-making.

The analysis showed that Anna regularly gathered and interpreted information about the students' learning needs and adapted feedback and learning activities to meet these needs. That is, she implemented a formative assessment practice that can be characterized as teacher-centered focusing on information gathered by the teacher (and not by the students). Anna described that she now gathers information about: students' prior knowledge and knowledge gained during lessons. These are presented below.

Gathering Information About Students' Prior Knowledge

Anna described that her notion that some of her students are likely to have insufficient prior knowledge to fully understand the course made her change her way of introducing new courses. Now she always starts by gathering information about her students' prior knowledge. She uses that information to plan her forthcoming lessons but also to act in a timely manner during the lesson itself. She mainly uses Google forms because, besides gathering information from all students, it compiles and presents the results straight away. Anna explains:

> The digital tool is essential to be able to work with formative assessment. To gather information, using pen and paper and spend time compiling the answers would be too time consuming, and you would not be able to act during the lesson itself.

Anna described that she now collects information in two main ways. First, in the initial lesson she asks her students to rate their understanding of some main concepts in the subject matter domain in order to acquire indications of their familiarity with the learning content. For example, the students were to rate 19 concepts such as gender, intersectionality, social constructivism, socialization, and feminism, and for each concept they were to answer whether they "never heard of it," "recognize it but can't explain it," "can explain it a little," or have "a total understanding of it." Anna described that since she felt she needed more information about students' actual understanding of the concepts (not merely their rating); she decided to ask the students to write down explanations for some concepts of her choice at the end of such lessons. That information, measuring their understanding of these concepts could be described as more accurate and useful to make decisions about what to emphasize in her subsequent teaching. Second, as a complement to the information about the students' perceived understanding of the subject matter, Anna uses the digital tool to administer questions measuring the actual extent (not only the perceived extent) to which they already possess the knowledge to be learned in the teaching and learning unit. For example, she poses the following question to her students: "Which of the following (parliament, municipality, county council, market, or other) decides on the following?" followed a number of decisions made in society such as "It is forbidden to hit children in Sweden." (See **Table 1** for other examples). In these instances the students answered anonymously.

Directly after the students have answered her questions, in any of these two ways, Anna and her students look at the results provided by Google forms. Anna shares the diagrams that show the results on a group level with her students, and she clarifies the learning objectives of the teaching and learning unit. These results indicate to both Anna and her students, part of the students' current knowledge in relation to the learning goals. Anna points out the challenge when it turns out that her students' prior knowledge differs considerably, by stating

> There are situations when some students can't identify the European countries using a map (that should have been learnt in middle school) when other students can account for the social, economic, political and cultural differences between Greece and Germany. . . . so where do I start, at middle school level, to include students with insufficient prior knowledge or should I start at the level where the students are expected to be?

Anna explained that she decided to aim to adjust her teaching on an individual level and, to be able to do so, she has changed her approach to have the students answering anonymously, as Anna said: "Now I often invite students to write their name, so that I, in peace and quiet after the lesson, can identify students who need extra support." Anna described that she now, knowing who they are, actively approaches these students during lessons to provide support, even if they do not ask for help.

At the end of the teaching and learning unit (for example, one month later), she then sometimes again lets the students rate their understanding of subject matter concepts in order to support their awareness of their learning progress.

Decisions, knowledge and skills

To gather information about students' prior knowledge entails some consecutive decisions. She needs to decide on what prior knowledge is needed and what she can expect her students to know when the students enroll for the course. Then she can decide on, for example what 19 concepts she should ask her students to rate and describe. That is, concepts that her students could be expected to already know, together with concepts that are likely to be new. To make these decisions she needs to have extensive knowledge of the subject matter and sufficient knowledge about learning goals from her students' prior courses.

Then, having information about students' prior knowledge puts additional demands on her decision making. Based on that information, Anna decides how to adjust her teaching to fit the majority of her students. However, when Anna found that group

level information is not enough to identify students who really need help (and may not ask for it) she decided to list the students' names as well. To support these identified students she made the decision to approach these students intentionally during lessons even when they don't ask for help.

Besides using gathered information to make decisions on planning future lessons, Anna takes the opportunity to provide timely feedback or instructions directly after the students answer her questions. The latter being a complex matter of instant decision making on what concepts to explain, what misunderstanding to challenge, what action will benefit students with insufficient prior-knowledge the most, and so forth. This kind of decision-making puts great demands on her skills to quickly assess and choose information about her students' shortcomings and provide feedback accordingly. Anna pointed out: "These decisions are made 'on the fly'. However, I have been teaching for a fairly long time and have some experience to rely on. I mean, I know what students usually find difficult."

Gathering Information About Students' Gained Knowledge

Anna reported that she now uses questions during and at the end of lessons to gather the information about what her students have learned during the lesson. Based on this information she decides what feedback to give the whole class or individual students, whether to focus more or less on certain content, and which learning activities would meet the class's or individual student's identified learning needs.

Anna furthermore describes how she tells her students that they will be requested to answer some questions after a learning activity. She thinks that if the students know in advance that they are expected to answer questions, they are given extra incentive to pay attention and to engage in their own learning; as Anna said:

> At the beginning of a lesson, when I am going to give a lecture or show an educational movie . . . I tell my students that I am going to give them questions using Google forms afterwards, and that we will discuss them. This is a way of making them more focused, paying attention and providing them with an opportunity, and for me, to check whether they understand the important stuff. . . . if not they can ask me or, when I get the information that something was really tricky, I know what I need to explain again or I let them practice more.

The students' answers to the questions, and their utterances in the discussion, provide Anna with information about the students' understanding of the learning content included in the lecture or in the movie. Based on her interpretation of this information, she makes decisions on which parts of the content the students have not yet grasped and therefore need immediate further clarification or attention during the next lesson.

When the students work with tasks, Anna walks around in the classroom to help her students. As described above, Anna also did this before the implementation of her formative assessment practice, but she has now changed her way of providing support. Instead of immediately helping her students, she now first requires them to orally formulate what they have understood and what exactly their problem is. She then interprets their formulations and asks them to respond to her interpretation. She said that the decision to change her responses in this way is based on her belief that it would increase the validity of her interpretations of what the students have understood so far as well as their learning needs, which in turn provides her with a better foundation for her decisions on what feedback would be most beneficial for the students' learning.

At the end of many lessons, Anna gives the students questions using Google forms in order to gather information about what they have learned from the lessons so far in the teaching and learning unit. For example, at the end of one lesson she returned to some of the concepts for which the students had rated their understanding earlier (e.g., social constructivism and feminism), and the students were now asked to "formulate a few sentences that show your understanding of these concepts." In these cases the students also provided their names together with their answers. Based on her interpretation of the students' learning needs, she then makes decisions about how to best support the students' learning in the following lesson. Generally, when she judges that many students lack sufficient understanding, she will revisit the content with the whole class during the next lesson; if on the other hand only a few students lack sufficient understanding, she will work with them separately the following lesson.

Anna furthermore points out that there are situations when the information from the student is insufficient to even try to understand their difficulty and other supporting strategies are needed; Anna describes:

> For example, students that are convinced that they will fail. Their answers don't entail any information besides "I don't know" and they do not seem to make an effort during the lessons. I have tried many different strategies to motivate them more, but the one that has been most successful is to divide the assignment into smaller and more defined parts. That will make them take one step at the time and I can provide timely and frequent feedback. This will make them feel competent, that they are able to complete one (or several) sub-tasks within a lesson.

Introducing subtasks to bring the students to initiate their work at all will create further possibilities for Anna to gain information about their learning needs. Anna pointed out her aim to prevent students from falling behind, and that besides making sure that all students really understand the key concept in the course, to actively approach students who have difficulties. This way of breaking down assignments into sub-tasks to overcome one difficulty at a time works for some of her other students as well. That is, if the assignment is to examine the political and cultural differences between Greece and Germany the first easy-solved sub-task could be to learn where these countries are on a map.

Decisions, knowledge and skills

Anna's decision to continuously gather and act on information from all students put great demands on her decision-making, several of which have been accounted for earlier (see *Gathering Information About Students' Prior Knowledge* section). Together with her decision to approach students whom she has identified as having difficulties (besides those who ask for help), she decided to base these interactions on formative assessment. That is, to gather information about and identify the difficulty before providing feedback. Furthermore, when Anna encounters students not active during lessons and unwilling to share their difficulties, Anna has gone through a series of decisions about trying out, evaluating and discharging supporting strategies. Her latest decision however, that of dividing and concretize assignments into sub-tasks, managed to bring these students to engage and feel competent in finalizing tasks during the lesson.

Summary

Anna's shift in focus from students' learning outcome at the end of a course to her students' learning process made her implement a formative assessment practice. The progress of her assessment practice could be described as: Moving from merely summative assessment, to adding formative assessment at a group level and thereafter also adding formative assessment at the individual level. Thus, she added a formative aim to her assessment practice, resulting in additional requirements on her decision-making and her knowledge and skills. She shifted from mainly eliciting information about her students' learning for grading purposes to using this information to adjust her teaching to fit her students' level of prior knowledge and to support them to attain the learning goals. This additional aim required her to make decisions she did not have to make before, such as deciding what information would be useful for making instructional adjustments, when and how this information should be collected, and how to act on this information to support her students' learning. For example, to gain extensive insight into her students' prior knowledge and learning achievements during lessons, and the heterogeneity thereof, she had to make a series of decisions. She needed to decide what prior knowledge of subject matter concepts was important for the students to have for the learning practice to be as efficient as possible, and thereafter decide how to design introductory lessons to target the students' lack of such knowledge. She had to decide when to intentionally approach students she identified as having specific learning needs, to provide repetitive instructions to smaller groups of students identified as falling behind and divide tasks into subtasks to fit students with low motivation and self-esteem, and so forth. But she also had to decide how to give students opportunities to choose tasks based on personal interests in order to give them the independence they needed to aim for course content that suited their level of knowledge. These decisions require teacher knowledge and skills that go beyond familiarity of national standards and curriculums. For example, Anna needed to gain insights into what prior subject matter knowledge the students could be expected, and would be necessary, to have when they enrolled in her class. She furthermore needs skills to choose, elicit, interpret and act on information about students' learning needs. But moreover, these skills included how to interpret and act instantly to be able to provide timely feedback during the lessons. What is noteworthy is that Anna realizes situations where the information about her students' knowledge and skills is insufficient and she needs to resort to supporting strategies, for example the design of sub-tasks, in the formative assessment process. Thus, the aim of using assessment for instructional purposes adds requirements of constant flexibility and choosing or inventing strategies.

Changes in the Students' Behavioral Engagement

Students in the intervention class increased their behavioral engagement between spring and fall, and had a more positive change than all of the comparison classes (**Table 2**). Three of the comparison classes actually show a decrease in behavioral engagement. The size of the change in students' behavioral engagement, as estimated by comparing the difference between fall and spring in the intervention class with the difference in each of the comparison classes, was between small and medium (from 0.24 to 0.64).

Changes in the Students' Type of Motivation

Table 3 shows that students in the intervention class increased their autonomous motivation between spring and fall, and had a more positive change than all of the comparison classes. The size of the change in students' autonomous motivation, as estimated by comparing the difference between fall and spring in the intervention class with the difference in each of the comparison classes, was close to medium for all comparisons (from 0.42 to 0.50).

Table 4 shows that students in the intervention class increased also their controlled motivation between spring and fall more than all of the comparison classes. The size of the change in students' controlled motivation, as estimated by comparing the difference between fall and spring in the intervention class with the difference in each of the comparison classes, was between small and large (from 0.28 to 0.74).

DISCUSSION

Characteristics of the Formative Assessment Practice and the Requirements on the Teacher's Decision-Making

The implementation of the formative assessment practice had a profound influence on the decisions about teaching and learning that Anna had to make. The analysis of Anna's implemented practice shows how such teacher-centered formative assessment put further demands on teacher decision-making than more traditional teaching practices. In both of these types of teaching practices teachers need to make decisions about how to present content, which tasks to use in learning activities and summative tests, what kind of feedback to give, and which grades to assign to students. However, in many traditional forms of

TABLE 2 | Behavioral engagement in the intervention and comparison classes in the fall and spring and their difference.

		Fall		Spring		Difference		
Class	**N**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**g[a]**
ChReA (Anna) (Intervention)	12	4.47	1.03	5.00	1.13	0.53	1.54	
BuCo	10	3.77	1.33	3.60	1.18	–0.17	0.54	0.58
ChReB	9	4.07	1.49	4.31	1.48	0.24	0.48	0.24
InTe	8	4.38	0.76	4.28	0.86	–0.10	0.61	0.50
SoScA	24	4.86	1.01	4.96	0.96	0.10	0.72	0.41
SoScB	21	4.56	1.12	4.38	1.15	–0.18	0.78	0.64

[a]Hedge's g for the difference in mean change when comparing the intervention class with the comparison class. Positive values mean that the intervention class had a more positive change compared to the comparison class.
Note: BuCo is the class from the Building and Construction Program, InTe is the class from the Industrial Technology Program, ChReA and ChReB are the two classes from the Child and Recreation Program (ChReA is the intervention class), and SoScA and SoScB are the two classes from the Social Science Program.

TABLE 3 | Autonomous motivation in the intervention and comparison classes in the fall and spring and their difference.

		Fall		Spring		Difference		
Class	**N**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**g[a]**
ChReA (Anna) (Intervention)	12	4.58	1.57	5.13	0.92	0.54	1.09	
BuCo	10	3.25	1.41	3.28	1.20	0.03	1.29	0.43
ChReB	9	3.96	1.33	4.05	1.32	0.09	0.55	0.50
InTe	8	4.08	1.13	4.25	1.22	0.17	0.40	0.42
SoScA	24	5.01	1.06	5.15	1.07	0.14	0.79	0.45
SoScB	21	4.08	1.01	4.06	1.43	–0.02	1.26	0.47

[a]Hedge's g for the difference in mean change when comparing the intervention class with the comparison class. Positive values mean that the intervention class had a more positive change compared to the comparison class.
Note: BuCo is the class from the Building and Construction Program, InTe is the class from the Industrial Technology Program, ChReA and ChReB are the two classes from the Child and Recreation Program (ChReA is the intervention class), and SoScA and SoScB are the two classes from the Social Science Program.

TABLE 4 | Controlled motivation in the intervention and comparison classes in the fall and spring and their difference.

		Fall		Spring		Difference		
Class	**N**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**g[a]**
ChReA (Anna) (Intervention)	12	4.35	1.19	4.85	1.15	0.50	1.17	
BuCo	10	4.32	0.86	4.52	0.67	0.20	0.93	0.28
ChReB	9	4.00	0.96	3.87	0.90	–0.13	0.99	0.57
InTe	8	4.06	0.77	3.85	0.76	–0.21	0.49	0.74
SaScA	24	4.76	1.40	4.69	1.21	–0.06	0.85	0.58
SaScB	21	5.09	1.03	5.10	1.11	0.01	1.25	0.40

[a]Hedge's g for the difference in mean change when comparing the intervention class with the comparison class. Positive values mean that the intervention class had a more positive change compared to the comparison class.
Note: BuCo is the class from the Building and Construction Program, InTe is the class from the Industrial Technology Program, ChReA and ChReB are the two classes from the Child and Recreation Program (ChReA is the intervention class), and SoScA and SoScB are the two classes from the Social Science Program.

teaching information about students' learning based on continuous assessments is not the focus when deciding on how to resolve pedagogical issues (Lloyd, 2019) or when monitoring the classroom practice (Shavelson and Stern, 1981). In contrast, the focus in Anna's formative assessment practice is on making pedagogical decisions based on continuously gathered empirical evidence about her students' learning. As a consequence, in line with arguments from several researchers (Brookhart, 2011; Means et al., 2011; Gummer and Mandinach 2015; Datnow and Hubbard, 2016) and empirically shown in the present study, the teacher in such a formative assessment practice also needs to make decisions about how to gather information about the students' knowledge and skills during the teaching and learning units, what this information means in terms of learning needs, and how to use the conclusions about learning needs to adapt feedback and learning activities to these needs. However, formative assessment may be carried out in different ways (Stobart and Hopfenbeck, 2014). Some formative assessment practices may have a positive effect on students' motivation while others may not, and the requirements of these different practices on teachers' decision making may not be the same. The present study exemplifies some of the decisions,

and the skills used to make them, that are required in a formative assessment practice in which both students' autonomous motivation and their behavioral engagement in learning activities increased. Studies that investigate effects of formative assessment on motivation within an ecologically valid, regular classroom environment are scarce (Hondrich et al., 2018), and we have not found any studies that have examined the requirements on teachers' decision making in formative assessment practices that have been empirically shown to have an impact on students' engagement or on autonomous and controlled motivation.

It should be noted that the formative assessment practice requires Anna to make some of the decisions under difficult conditions, and her disposition and skills need to afford her the ability to cope with making decisions under such conditions. These conditions are in many ways more difficult than those in, for example, practices in which the formative aspect of the practice only is constituted by formative use of summative assessment data. Instructional decisions based on summative assessment data are made much more infrequent and under less time pressure. That kind of data does not appear to inform the instructional decisions in the day-to-day practice (Oláh et al., 2010; Hoover and Abrams, 2013). In order to be able to adapt the teaching during a lesson, Anna needs to be able to develop or choose tasks that provide information about students' conceptual understanding but do not take a long time for the students to answer and for Anna to assess. Moreover, because the formative assessment practice is founded on the idea of continuously adapting teaching and learning to all students' learning needs, it is not sufficient to gather information only about a few students' learning needs or to only adapt the teaching in coming lessons. Therefore, Anna needs to be able to administer the questions and collect and interpret the answers from all students even in the middle of lessons. Letting the individual students who raise their hands answer the questions would not suffice, and the use of an all-response system such as Google Forms allows her to see the responses from all students at the same time. When adaptations of teaching are made during the same lesson that the assessment is done, Anna needs to make decisions both under time pressure and without knowing in advance exactly which learning needs the assessment will show. In her formative assessment practice, Anna will much more often than in her previous more traditional way of teaching make decisions on how to use the conclusions about all the students' learning needs. This means that she much more often is required to make decisions about how to adapt teaching to a class of students that may have different learning needs and must be able to individualize instruction and learning activities to these different needs when her interpretation of the assessment information suggests this to be most useful for the students' learning. Whatever actions are taken, the decisions about actions need to be taken based on the identified learning needs and not on a predetermined plan for the teaching and learning unit. The latter, in contrast, would generally be a cornerstone of a more traditional teaching practice (Shavelson and Stern, 1981).

It should be noted that the additional decisions teachers need to make, and the skills required to make them, in the formative assessment practice in comparison with a more traditional way of teaching are by no means trivial. Thus, as is argued by, for example, Mandinach and Gummer (2016), in order to provide teachers with reasonable possibilities to implement this kind of practice it would be important for teacher education and professional development programs to take into account the decisions and skills required to carry out this practice. In the present study we have identified some of the skills that may be useful to take into account when supporting pre- and in-service teachers in developing the skills necessary for implementing formative assessment that have a positive effect on motivation. For example, in line with the results of this study, our practical experience suggests that it may be crucial that professional development programs help teachers in how to use assessment information to adapt their teaching to their students' often different learning needs. In addition, the teachers may need assistance in finding ways to carry out formative assessment practices in the practicalities of disorderly classroom situations. To accomplish such assistance, professional development leaders may also need to collect evidence of the teachers' difficulties and successes in the actual flow of their classroom activities to be able to provide sufficient assistance.

Changes in the Students' Behavioral Engagement and Type of Motivation

However, using these additionally required skills in making the decisions and implementing this practice may pay off in terms of positive student outcomes. The students' behavioral engagement and autonomous motivation increased in the intervention class both in absolute numbers and compared to all of the comparison classes. The changes compared to the comparison classes were mostly of medium size. Thus, the change in students' autonomous motivation in the present study was higher than the changes in autonomous motivation coming from formative assessment implementations of teachers who did not receive comprehensive professional development support (e.g., Förster and Souvignier, 2014), and from interventions in which teachers were provided with a short professional development course (Hondrich et al., 2018) or a digital formative assessment tool (Faber et al., 2017) to aid the formative assessment processes of providing student assignments and feedback. The change in autonomous motivation was of a similar order of magnitude as when teachers were provided with information about how to implement a formative assessment practice that involved both teachers and students in the core processes of formative assessment (Meusen-Beekman et al., 2016). The change in students' engagement were of a similar order of magnitude as when a researcher taught self-assessment strategies in a student-centered formative assessment practice (Wong, 2017).

In this study we also investigated the change in students' controlled forms of motivation. This is not commonly done in existing studies of effects of formative assessment on students' motivation. Interestingly, the results show that not only autonomous forms of motivation increased more in the intervention class than in the comparison classes. Controlled forms of motivation also increased in the intervention class both in absolute numbers and compared to all of the comparison classes.

In comparison with the comparison classes, these students experienced both the autonomous reasons and the controlled reasons for engaging in learning to be more important after the formative assessment intervention than before. As a consequence, there was no shift away from more controlled forms of motivation toward more autonomous forms of motivation among the students. This shows the value of investigating changes of different types of motivation, not just of autonomous motivation. Any type of motivation may enhance students' engagement in learning activities, but because autonomous motivation has been associated with more positive emotions and better learning strategies than controlled motivation (Ryan and Deci, 2000), it might have been even more valuable for the students if the increase in behavioral engagement and autonomous motivation had been achieved without the corresponding increase in controlled motivation.

The present study does not investigate the reasons for the change in students' motivation. But the characteristics of the practice provide some indications of possible reasons. Anna began to require her students to orally formulate what they had understood and what exactly they perceived their problem to be before she provided them with help, she also started to using google forms which required all of her students (and not only a few students) to respond to her questions, and sometimes she informed her students that after a presentation of content or some other activity they would be given questions about the content. These activities may have given the students direct incitement to engage in learning during these occasions, which may have affected their learning habits in general toward more engagement also in other learning activities. Anna's more frequent assessments of her students' knowledge and skills followed by feedback and learning activities adapted to the information from the assessments, may have helped the students to acknowledge that they have learned and can meet goals and criteria. The feedback and learning activities adapted to information about students' learning needs may also have increased students' actual learning. In line with theorizing by for example Heritage and Wylie (2018), these experiences may have facilitated students' development of an identity as effective and capable learners, and as a consequence enhanced students' motivation to engage in learning activities. In line with Shepard et al. (2018) theorizing, Anna's formative assessment practices in which feedback helps students see what they have learned and how to improve may also have fostered a learning orientation in which students find it personally valuable to engage in learning activities and thus feel more autonomous in their motivation. Finally, in line with arguments by Hondrich et al. (2018) and Pat-El et al. (2012), Anna's focus on gathering information about students' knowledge and skills and providing feedback that both helps students monitor their learning progress and provides support for how goals and criteria can be met, may have enhanced students' satisfaction of the psychological need for competence, which according to self-determination theory (Ryan and Deci, 2000) influences students' autonomous forms of motivation.

Limitations of the Present Study and Possible Future Studies

The formative assessment practice described in the present study is teacher-centered in the sense that the teacher is the main active agent in the core formative assessment processes. Formative assessment may also have other foci. For example, formative assessment may combine the characteristics of a teacher-centered approach with practice in which the students are more proactive in the formative assessment processes. In such practices, the students would also be engaged in peer and self-assessment followed by adapting feedback and learning based on the identified learning needs. The teacher's role is to support the students in these processes. This approach to formative assessment would require the teacher to be involved in even more types of decision making about teaching and learning, and would require even more skills than the practice analyzed in the present study. Such practice may produce other effects on students' engagement and type of motivation. The shift from a practice in which the teacher is seen as the agent responsible for most decisions about teaching and learning to a practice in which the responsibility for these decisions are more balanced between the teacher and the students might cause an increase in students' engagement in learning activities (Brookhart, 2013; Heritage and Wylie, 2018), and in autonomous motivation without a similar increase in controlled motivation (Shepard et al., 2018). Future studies investigating this hypothesis would be a valuable contribution to research on the effects of formative assessment on motivation.

One limitation of this study is that only one teacher's implementation of formative assessment was investigated. This is sufficient for identifying some of the decisions required to be made in teacher-centered formative assessment practices and the skills needed to make them. However, in the investigation of the changes in students' motivation, this opens up for some uncertainties about whether there are other characteristics of the classroom practice than formative assessment that may have contributed to the positive changes in the students' motivation. In addition, only having one intervention class also makes the study underpowered, which means that changes that are not very large will not be detected in significance analyses. This also makes it uncertain as to whether the results would be similar with other students and in other contexts. A second limitation of the study is that there is no analysis of the classroom practices in the comparison groups.

However, to avoid the risk of different changes in the intervention group and in the comparison groups on the outcome variables (behavioral engagement, and autonomous and controlled forms of motivation) not being due to the implemented formative assessment practice but to differences in prior academic achievement, the comparison classes were chosen based on the fact that classes in these programs, despite program differences, did not differ much regarding prior academic achievement. Furthermore, we have used both prequestionnaires and postquestionnaires to measure the changes on the outcome variables. In this way, the risk that students' prior forms of motivation and behavioral engagement would influence the changes on the outcome variables is minimized.

Another possible threat to the validity of a conclusion that the change in students' motivation is due to the formative assessment practice would be if those students in the intervention class who increased their engagement and motivation the most chose to

participate in the questionnaire survey to a higher extent than other students, and if the opposite was true for the students in the comparison classes. However, since only three persons declined to participate, and they were spread over the classes, almost all of the non-participating students were those who happened to not be present on both occasions when the questionnaire was administered. Such non-participation could affect the mean values of the students' answers on each questionnaire item because students who attend most classes might be overrepresented in our samples. However, such overrepresentation would be similarly distributed over all classes, and thus not affect the results of the study.

Another variable that could have had an influence on the results are the teaching practices in the comparison classes. If some of the comparison teachers also would have implemented formative assessment practices, it would be difficult to draw any conclusions about the higher increase in motivation in the intervention class being due to the implemented formative assessment practice. However, none of the comparison teachers had participated in any professional development program in formative assessment, and they continued to teach in the ways they had taught before. This makes it highly unlikely that they would have engaged in formative assessment practices. Furthermore, the results show that the intervention group increased more than all of the comparison classes on all outcome variables. Thus, whatever characteristics of the teaching in the comparison classes, none of them had the same influence on the outcome variables as the intervention teacher's implemented formative assessment practice.

Another possibility is that the intervention teacher was especially proficient in enhancing students' motivation in other ways than by the use of formative assessment, and that those ways are the reasons for the changes in motivation being more positive in the intervention class than in the comparison classes. This cannot be ruled out but may be less likely since the intervention teacher was not selected to the study for any other reason than that she had participated in a professional development program in formative assessment to which she had not volunteered and was not selected based on any of her characteristics. She came from another school in which all teachers participated in that professional development program, so the reason for her participation was just that she happened to be at that school when the program was carried out.

Hence, the evidence supporting the conclusion that the implemented formative assessment practice is the reason for the increase in the students' motivation being larger in the intervention class than in the comparison classes seems to be much stronger than the evidence supporting other possible conclusions. However, future studies using larger samples of intervention teachers and involving more thorough analyses of the classroom practices in the comparison groups would be valuable to be able to make more generalizable conclusions about the effects of formative assessment practices on students' motivation both in terms of their type of motivation and their engagement in learning activities.

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 participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors have contributed to the introduction part and formulation of research questions as well to the discussion. GN, CA and CG have contributed to qualitative data collection and that part of the Results. TP and BP have contributed to the quantitative data collection and that part of the Results.

REFERENCES

Baas, D., Vermeulen, M., Castelijns, J., Martens, R., and Segers, M. (2019). Portfolios as a Tool for AfL and Student Motivation: Are They Related? *Assess. Educ. Principles, Pol. Pract.* 27, 444–462. doi:10.1080/0969594X.2019.1653824

Bishop, A. J. (1976). Decision-making, the Intervening Variable. *Educ. Stud. Math.* 7 (1/2), 41–47. doi:10.1007/bf00144357

Black, P., and Wiliam, D. (2009). Developing the Theory of Formative Assessment. *Educ. Asse Eval. Acc.* 21 (1), 5–31. doi:10.1007/s11092-008-9068-5

Borko, H., Roberts, S. A., Shavelson, R., Clarkson, P., and Presmeg, N. (2008). "Teachers' Decision Making: from Alan J. Bishop to Today," in *Critical Issues in Mathematics Education* (Boston, MA: Springer), 37–67. doi:10.1007/978-0-387-09673-5_4

Brookhart, S. M. (2013). "Classroom Assessment in the Context of Motivation Theory and Research," in *SAGE Handbook of Research on Classroom Assessment*. Editor J. H. McMillan (Thousand Oaks, CA: SAGE), 33–54.

Brookhart, S. M. (2011). Educational Assessment Knowledge and Skills for Teachers. *Educ. Meas. Issues Pract.* 30 (1), 3–12. doi:10.1111/j.1745-3992.2010.00195.x

Clark, I. (2012). Formative Assessment: Assessment Is for Self-Regulated Learning. *Educ. Psychol. Rev.* 24 (2), 205–249. doi:10.1007/s10648-011-9191-6

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd edition. New York, NY: Routledge.

Costello, A. B., and Osborne, J. (2005). Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most from Your Analysis. *Pract. Assess. Res. Eval.* 10 (1), 7. doi:10.7275/jyj1-4868

Datnow, A., and Hubbard, L. (2016). Teacher Capacity for and Beliefs about Data-Driven Decision Making: A Literature Review of International Research. *J. Educ. Change* 17 (1), 7–28. doi:10.1007/s10833-015-9264-2

Datnow, A., and Hubbard, L. (2015). Teachers' Use of Data to Inform Instruction: Lessons from the Past and Prospects for the Future. *Teach. Coll. Rec.* 117 (4), 1–26. doi:10.1007/s10833-015-9264-2

Faber, J. M., Luyten, H., and Visscher, A. J. (2017). The Effects of a Digital Formative Assessment Tool on Mathematics Achievement and Student

Motivation: Results of a Randomized Experiment. *Comput. Edu.* 106, 83–96. doi:10.1016/j.compedu.2016.12.001

Federici, R. A., Caspersen, J., and Wendelborg, C. (2016). Students' Perceptions of Teacher Support, Numeracy, and Assessment for Learning: Relations with Motivational Responses and Mastery Experiences. *Ies* 9 (10), 1–15. doi:10.5539/ies.v9n10p1

Förster, N., and Souvignier, E. (2014). Learning Progress Assessment and Goal Setting: Effects on Reading Achievement, Reading Motivation and Reading Self-Concept. *Learn. Instruction* 32, 91–100. doi:10.1016/j.learninstruc.2014.02.002

Fredricks, J. A., Blumenfeld, P. C., and Paris, A. H. (2004). School Engagement: Potential of the Concept, State of the Evidence. *Rev. Educ. Res.* 74 (1), 59–109. doi:10.3102/00346543074001059

Gan, Z., He, J., He, J., and Liu, F. (2019). Understanding Classroom Assessment Practices and Learning Motivation in Secondary EFL Students. *J. Asia TEFL* 16 (3), 783–800. doi:10.18823/asiatefl.2019.16.3.2.783

Ghaffar, M. A., Khairallah, M., and Salloum, S. (2020). Co-constructed Rubrics and Assessment for Learning: The Impact on Middle School Students' Attitudes and Writing Skills. *Assessing Writing* 45, 100468. doi:10.1016/j.asw.2020.100468

Gummer, E., and Mandinach, E. (2015). Building a Conceptual Framework for Data Literacy. *Teach. Coll. Rec.* 117 (4), 1–22.

Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E., Supovitz, J., and Wayman, J. (2009). IES Practice Guide: Using Student Achievement Data to Support Instructional Decision Making *(NCEE 2009-4067)*. Washington, DC: National Center for Education Evaluation and Regional Assistance.

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *J. Educ. Stat.* 6 (2), 107–128. doi:10.3102/10769986006002107

Heritage, M., and Wylie, C. (2018). Reaping the Benefits of Assessment for Learning: Achievement, Identity, and Equity. *ZDM Math. Edu.* 50 (4), 729–741. doi:10.1007/s11858-018-0943-3

Hondrich, A. L., Decristan, J., Hertel, S., and Klieme, E. (2018). Formative Assessment and Intrinsic Motivation: The Mediating Role of Perceived Competence. *Z. Erziehungswiss* 21 (4), 717–734. doi:10.1007/s11618-018-0833-z

Hoover, N. R., and Abrams, L. M. (2013). Teachers' Instructional Use of Summative Student Assessment Data. *Appl. Meas. Edu.* 26 (3), 219–231. doi:10.1080/08957347.2013.793187

Lloyd, C. A. (2019). Exploring the Real-World Decision-Making of Novice and Experienced Teachers. *J. Further Higher Edu.* 43 (2), 1–17. doi:10.1080/0309877x.2017.1357070

Mandinach, E. B., and Gummer, E. S. (2016). What Does it Mean for Teachers to Be Data Literate: Laying Out the Skills, Knowledge, and Dispositions. *Teach. Teach. Edu.* 60, 366–376. doi:10.1016/j.tate.2016.07.011

Mandinach, E., and Schildkamp, K. (2020). Misconceptions about Data-Based Decision Making in Education: An Exploration of the Literature. *Stud. Educ. Eval.* 23, 100842. doi:10.1016/j.stueduc.2020.100842

Means, B., Chen, E., DeBarger, A., and Padilla, C. (2011). *Teachers' Ability to Use Data to Inform Instruction: Challenges and Supports.* Washington, DC: US Department of Education, Office of Planning, Evaluation, and Policy Development.

Meijer, P. C., Verloop, N., and Beijaard, D. (2002). Multi-method Triangulation in a Qualitative Study on Teachers' Practical Knowledge: An Attempt to Increase Internal Validity. *Qual. Quantity* 36 (2), 145–167. doi:10.1023/a:1014984232147

Meusen-Beekman, K. D., Joosten-ten Brinke, D., and Boshuizen, H. P. A. (2016). Effects of Formative Assessments to Develop Self-Regulation Among Sixth Grade Students: Results from a Randomized Controlled Intervention. *Stud. Educ. Eval.* 51, 126–136. doi:10.1016/j.stueduc.2016.10.008

Oláh, L. N., Lawrence, N. R., and Riggan, M. (2010). Learning to Learn from Benchmark Assessment Data: How Teachers Analyze Results. *Peabody J. Edu.* 85 (2), 226–245. doi:10.1080/01619561003688688

Pat-El, R., Tillema, H., and van Koppen, S. W. M. (2012). Effects of Formative Feedback on Intrinsic Motivation: Examining Ethnic Differences. *Learn. Individual Differences* 22 (4), 449–454. doi:10.1016/j.lindif.2012.04.001

Ryan, R. M., and Connell, J. P. (1989). Perceived Locus of Causality and Internalization: Examining Reasons for Acting in Two Domains. *J. Personal. Soc. Psychol.* 57 (5), 749–761. doi:10.1037/0022-3514.57.5.749

Ryan, R. M., and Deci, E. L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemp. Educ. Psychol.* 25 (1), 54–67. doi:10.1006/ceps.1999.1020

Schneider, M. C., and Randel, B. (2010). "Research on Characteristics of Effective Professional Development Programs for Enhancing Educators' Skills in Formative Assessment," in *Handbook of Formative Assessment*. Editors H. L. Andrade and G. J. Cizek (Abingdon: Routledge), 251–276.

Schoenfeld, A. (1998). Toward a Theory of Teaching-In-Context. *Issues Edu.* 4 (1), 1–94. doi:10.1016/s1080-9724(99)80076-7

Shavelson, R. J., and Stern, P. (1981). Research on Teachers' Pedagogical Thoughts, Judgments, Decisions, and Behavior. *Rev. Educ. Res.* 51, 455–498. doi:10.3102/00346543051004455

Shepard, L. A., Penuel, W. R., and Pellegrino, J. W. (2018). Using Learning and Motivation Theories to Coherently Link Formative Assessment, Grading Practices, and Large-Scale Assessment. *Educ. Meas. Issues Pract.* 37 (1), 21–34. doi:10.1111/emip.12189

Skinner, E. A., Kindermann, T. A., Connell, J. P., and Wellborn, J. G. (2009). "Engagement and Disaffection as Organizational Constructs in the Dynamics of Motivational Development," in *Handbook of Motivation in School*. Editors K. Wentzel and A. Wigfield (New York: Routledge), 223–245.

Stobart, G., and Hopfenbeck, T. (2014). "Assessment for Learning and Formative Assessment," in *State of the Field Review: Assessment and Learning*. Editors J. Baird, T. Hopfenbeck, P. Newton, G. Stobart, and A. Steen-Utheim (Oslo, Norway: Norwegian Knowledge Centre for Education), 30–50.

Wiliam, D., and Thompson, M. (2008). "Integrating Assessment with Learning: What Will it Take to Make it Work?," in *The Future of Assessment: Shaping Teaching and Learning*. Editor C. A. Dwyer (Mahwah, NJ: Lawrence Erlbaum Associates), 53–82.

Winberg, T. M., Hofverberg, A., and Lindfors, M. (2019). Relationships between Epistemic Beliefs and Achievement Goals: Developmental Trends over Grades 5-11. *Eur. J. Psychol. Educ.* 34 (2), 295–315. doi:10.1007/s10212-018-0391-z

Wong, H. M. (2017). Implementing Self-Assessment in Singapore Primary Schools: Effects on Students' Perceptions of Self-Assessment. *Pedagogies: Int. J.* 12 (4), 391–409. doi:10.1080/1554480X.2017.1362348

Zhang, W. (2017). Using Classroom Assessment to Promote Self-Regulated Learning and the Factors Influencing its (In)effectiveness. *Front. Educ. China* 12 (2), 261–295. doi:10.1007/s11516-017-0019-0

Zimmerman, B. J. (2000). "Attaining Self-Regulation," in *Handbook of Self-Regulation*. Editors M. Boekaerts, P. R. Pintrich, and M. Zeidner (San Diego: Academic Press), 13–39. doi:10.1016/b978-012109890-2/50031-7

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.

APPENDIX A

QUESTIONNAIRE ITEMS

The items measuring students' engagement in learning activities are statements that the students were asked to mark to what extent they agreed with on a scale from 1 (not at all) to 7 (fully agree). The items measuring students' type of motivation are statements of reasons for working during lessons or for learning the course content. The students were asked to mark to what extent these reasons were important on a scale from 1 (not at all a reason) to 7 (really important reason). The word "[subject]" was replaced with the particular school subject the students were studying, for example sociology or history.

Items measuring behavioral engagement

1. I am always focused on what I'm supposed to do during lessons.
2. I use all given time during lessons to work with [subject].
3. If I encounter something difficult during this course, I make a strong effort to try to understand.
4. During lessons, I do not think about anything other than what I am supposed to learn.
5. I always try to learn as much as possible in this course.

Items measuring autonomous motivation

1. When I work during lessons with the tasks I have been assigned, I do it because it's good for me.
2. When I work during lessons with the tasks I have been assigned, I do it because I want to learn new things.
3. When I work during lessons with the tasks I have been assigned, I do it because it's fun.
4. When I work during lessons with the tasks I have been assigned, I do it because I like it.
5. When I try to learn the content of this course, I do it because it's fun to learn new things.
6. When I try to learn the content of this course, I do it because it's interesting.

Items measuring controlled motivation

1. When I work during lessons with the tasks I have been assigned, I do that because I want the teacher to think that I am a good student.
2. When I work during lessons with the tasks I have been assigned, I do that because I will feel ashamed if I don't do them.
3. When I work during lessons with the tasks I have been assigned, I do that because the teacher says I should do it.
4. When I try to learn the content of this course, I do it because others think it is important that I get the best grades possible.
5. When I try to learn the content of this course, I do it because I will feel bad if I don't perform well.
6. When I try to learn the content of this course, I do it because it's expected of me.

frontiers in Education

ORIGINAL RESEARCH
published: 17 March 2022
doi: 10.3389/feduc.2022.830311

Inside Teacher Assessment Decision-Making: From Judgement Gestalts to Assessment Pathways

De Van Phung[1] and Michael Michell[2]*

[1] Resource Development Institute, Tra Vinh University, Tra Vinh, Vietnam, [2] School of Education, Faculty of Arts and Social Science, The University of New South Wales, Kensington, NSW, Australia

Assessment decision-making is an integral part of teacher practice. Issues related to its trustworthiness have always been a major area of concern, particularly variability and consistency of teacher judgment. While there has been extensive research on factors affecting variability, little is understood about the cognitive processes that work to improve the trustworthiness of assessment. Even in an educational system like Australia, where teacher-based assessment in mainstream classrooms is widespread, it has only been relatively recently that there have been initiatives to enhance the trustworthiness of teacher assessment of English as a second or additional language (EAL). To date, how teachers make their decisions in assessing student oral language development has not been well studied. This paper reports on the nature and dynamics of teacher decision-making as part of a larger study aimed at exploring variability of teacher-based assessment when using the oral assessment tasks and protocols developed as part of the Victorian project, Tools to Enhance Assessment Literacy for Teachers of English as an Additional Language (TEAL). Employing a mixed-method research approach, this study investigated the assessment judgements of 12 experienced NSW primary and secondary EAL teachers through survey, assessment activity, think-aloud protocols and individual follow-up interviews. The paper highlights the key role of teachers' first impressions, or judgement Gestalts, in forming holistic appraisals shaping subsequent assessment decision-making pathways. Based on the data, a model identifying three assessment decision-making pathways is proposed which provides a new lens for understanding differences in teachers' final assessment judgements of student oral language performances and their relative trustworthiness. Implications of the model for assessment theory and practice, teacher education, and future research are discussed.

Keywords: teacher decision-making, teacher-based assessment, language assessment, Gestalt, holistic and analytical assessment, appraisal, trustworthiness

OPEN ACCESS

Edited by:
Chris Davison, University of New South Wales, Australia

Reviewed by:
Yongcan Liu, University of Cambridge, United Kingdom
Wei Shin Leong, Ministry of Education, Singapore

****Correspondence:***
De Van Phung
dephung@tvu.edu.vn

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

Received: *07 December 2021*
Accepted: *20 January 2022*
Published: *17 March 2022*

Citation:
Phung DV and Michell M (2022) Inside Teacher Assessment Decision-Making: From Judgement Gestalts to Assessment Pathways. Front. Educ. 7:830311. doi: 10.3389/feduc.2022.830311

INTRODUCTION

Sound assessment decision-making underpins the trustworthiness of teacher-based assessment in both general and language teaching contexts. The trustworthiness of teacher-based language assessments has always been a matter of concern. Teachers' grading decisions (McMillan and Nash, 2000), inter- and intra-rater reliability (McNamara, 1996, 2000; Gamaroff, 2000) and language performance assessment are all subject to variability (Lado, 1961; Huot, 1990; Hamp-Lyons, 1991;

Janopoulos, 1993; Williamson and Huot, 1993; Weigle, 2002). Likewise, the inherent subjectivity of teacher-based assessment (McNamara, 1996, 2000) challenges the consistency (Luoma, 2004; Taylor, 2006) of teacher assessment decision-making. Moves towards assessment for learning as a trustworthy alternative to standardised testing (Stiggins, 2002; Smith, 2003; Davison, 2004, 2013; Popham, 2004, 2014; Davison and Leung, 2009) have only intensified the need to address these long-standing issues in classroom assessment (Anderson, 2003; Brookhart, 2003, 2011; McMillan, 2003; Harlen, 2005; Joughin, 2009a,b; Klenowski, 2013).

While there has been extensive research on factors affecting variability and consistency, teacher thinking processes affecting the trustworthiness of teacher-based assessment is little understood. Recent initiatives to enhance the trustworthiness of English language teacher assessment in Australia have focused on improving teachers' assessment literacy through collective socio-technical systems of support fostering moderated assessment practices around shared tools and resources (Davison, 2008, 2019, 2021; Michell and Davison, 2020). Over the last decade or so, the trustworthiness of teacher assessment judgements has been the central concern of assessment moderation studies (Klenowski et al., 2007; Klenowski and Wyatt-Smith, 2010; Wyatt-Smith et al., 2010; Adie et al., 2012; Wyatt-Smith and Klenowski, 2013, Wyatt-Smith and Klenowski, 2014) as well as individual teacher assessment studies (Crisp, 2010, 2013, 2017). This research on assessment judgement invites wider consideration of the nature of human judgement (Cooksey, 1996; Laming, 2004) and its operation as part of teacher cognition (Clark and Peterson, 1984; Freeman, 2002; Borg, 2009; Kubanyiova and Feryok, 2015) and classroom practice (Yin, 2010; Allal, 2013; Glogger-Frey et al., 2018).

Research on teacher assessment judgement has highlighted fundamental categories of holistic and analytical thinking and their interaction in assessment decision-making (e.g., Thomas, 1994; Anderson, 2003; Newton, 2007; Sadler, 2009; Crisp, 2017). These modes of thinking have a long history in psychological research. Kahneman's (2011) two modes of thinking: System 1—"fast," automatic, intuitive impressions and System 2—"slow" conscious, effortful attention—expand our understanding of holistic and analytical judgements and their respective and complementary operation in reliable decision-making and development of trustworthy expertise. It is in this context that Gestalt theory (Wagemans et al., 2012a,b; Wertheimer, 2012; Koffka, 2013) offers further insight into the holistic, impressionistic System 1 of language assessment. Although variability in classroom language assessment is an inherent characteristic of human assessment (McNamara, 1996; Davison, 2004; Davison and Leung, 2009), an understanding of the nature and dynamics of System 1 and 2 modes of thinking can be applied to enhancing the trustworthiness of teacher language assessment judgements and decision-making "from the inside."

In reviewing relevant research and reporting on the study findings, this paper outlines the following argument: (a) the situated cognitive processes underpinning teacher assessment practice is critical but is still underexplored; (b) such cognitive processes can be productively researched from the perspective of teacher judgement and decision-making; (c) holistic and analytical thinking and their dynamic interplay are fundamental thinking processes in how teachers form assessment judgements and decisions; (d) judgement Gestalts, conceptualised in this paper as holistic, intuitive assessment impressions, play a crucial role in teacher assessment shaping different assessment decision-making pathways; (e) a model making these assessment pathways and their contributory factors transparent can help teachers better understand their own assessment decision-making and ultimately improve the trustworthiness of teacher-based assessment and teacher assessment literacy.

Teacher Assessment Decision-Making

Assessment for Learning, the idea that assessment should be designed to promote student learning and thus be integrated with instruction (e.g., Black and Wiliam, 1998; Shepard, 2000, 2001; Stiggins, 2002; Stiggins et al., 2004; Gipps, 2012) "brings the teacher back in" (Michell and Davison, 2020) as leading agents of learning oriented assessment (Carless, 2007; Turner and Purpura, 2016). This renewed emphasis on formative, teacher-based classroom assessment has been accompanied by a paradigm shift in conceptions of assessment from *assessment-as-measurement* to *assessment-as-judgement*:

> as the role of assessment in learning has moved to the foreground of our thinking about assessment, a parallel shift has occurred towards the conception of assessment as the exercise of professional judgement and away from its conceptualisation as a form of quasi-measurement (Joughin, 2009a, p. 1 original italics).

As Sadler (2009) has noted, the traditional measurement model of assessment is reflected in the quantitative language of "gauging" the "extent of" learning, while the judgement model employs the qualitative language of "evaluation," "quality," and "judgement".

This shift has brought about a reconsideration of psychometric methods developed to ensure test validity and reliability and have lead to a reconception of what these standards look like in classrooms (e.g., Brookhart, 2003; Moss, 2003; Smith, 2003). Traditional standards of validity, reliability and fairness break down when applied to classroom assessment that support learning and new approaches to quality standards for assessment are required (Joughin, 2009a,b). Based on a critique that "measurement theory expects real and stable objects, not fluctuating and contested social constructs" (Knight, 2007, p. 77) of classrooms, some researchers have called for "classroometric" (Brookhart, 2003) or "edumetric" (Dochy, 2009) approaches to redesigning classroom assessment to meet the learning needs of students rather than satisfying the technical, psychometric properties of external testing. In this context, teachers' practical, pedagogical needs are foregrounded as necessary starting points for such designs (Davison and Michell, 2014) and issues of assessment validity and reliability are being reconsidered in terms of trustworthiness (Davison, 2004, 2017; Leung, 2013; Alonzo, 2019) and teacher assessment literacy (e.g., Mertler, 2004; Popham, 2004, 2009, 2014; Taylor, 2009; Brookhart, 2011; Koh, 2011; Xu and Brown, 2016; Davison, 2017).

The move to assessment-as-judgement highlights the evaluative nature of teacher assessment decision-making. Assessment judgements *are decisions* about the quality of

students' work and the best course of action the teacher might take in light of these decisions (e.g., Cooksey et al., 2007). Teacher-based assessment therefore brings to the fore considerations of the nature, development and exercise of human judgement in assessment, and these considerations are central to any theorising of assessment trustworthiness and teacher assessment literacy. Evaluative and inferential judgement is the epistemic core of teacher assessment decision-making:

> The act of assessment consists of appraising the quality of what students have done in response to a set task so that we can infer what students know (Sadler private communication quoted in Joughin, 2009b, p. 16, original italics)

Thus, judgement is appraisal—a decision concerning the value or quality of a performance or perceived competence which applies regardless of assessment purpose, participants or method. All judgements are, by nature, summative—even those made for formative purposes—there is no such thing as a formative judgement (Newton, 2007; Taras, 2009).

Underpinning this judgement-centred understanding of teacher assessment is the nature of teacher expertise that enables it. This expertise has been described as connoisseurship (Eisner, 1998)—a highly developed form of competence in qualitative appraisal, where "the expert is able to give a comprehensive, valid and carefully reasoned explanation for a *particular appraisal*, yet is unable to do so for the *general case*" (Sadler, 2009, p. 57, author italics). Teachers develop such expertise through extensive engagement and "reflection on action" in particular classroom events and situations. An implication of this is that models of teacher assessment decision-making that do not consider the exercise of professional judgement ignore the nature and role of language teacher cognition and epistemology (Borg, 2009; Kubanyiova and Feryok, 2015) in which teaching and assessment is grounded.

In classroom contexts, teacher assessment decision-making is a multi-step process in which teachers form judgements about the quality of student work or performance from available information and then relate these judgements as a score to a rubric, criteria, scale, standard or continuum. Sadler (1998) describes classroom assessment events as a common three stage structure of assessment judgement formation involving (1) teacher attention is drawn to student output, (2) teacher assessment of student output against some given scoring rubric and (3) teacher judgement or action decision. At each decision point in this process, different teachers tend to refer to and apply different resources to make their judgements. In assessing student task performance, teachers typically look first at student output information from different sources to gain an initial overall impression of students' abilities (Anderson, 2003; Crisp, 2017). During this stage, teachers rarely examine assessment rubrics or rating scales.

Within this process, two key modes of judgement are identified—holistic and analytical: "holistic grading involves appraising student works as integrated entities; analytic grading requires criterion-by-criterion judgements" (Sadler, 2009, p. 49). Newton (2007) describes these two judgment modes as being on a summative-descriptive continuum where *summative judgements* are characterised by appraisal—a decision concerning the value or quality of a performance or perceived competence and *descriptive judgements* are characterised by analysis—a reflection on the nature of the performance or perceived competence (p. 158).

Holistic assessment focuses on the overall quality of student work, rather than on its separate properties, and is foregrounded in both initial and final stages of the assessment process:

> In holistic, or global grading, the teacher responds to a student's work as a whole, then directly maps its quality to a notional point on the grade scale. Although the teacher may note specific features that stand out while appraising, arriving directly at a global judgement is foremost. Reflection on that judgement gives rise to an explanation, which necessarily refers to criteria. Holistic grading is sometimes characterised as impressionistic or intuitive (Sadler, 2009, p. 46).

Holistic assessment in the form of overall teacher judgements (OTJ) were found to be both lynch-pin and Achilles' heel of New Zealand education reform. Teachers were required to draw on and synthesise multiple sources of assessment information to make overall judgements about students' performance against National Standards. The Standards, however, were broad multi-criteria descriptors identified by Sadler (1985) as "fuzzy" standards. The study found that teachers formed somewhat equivocal overall judgements against the standards in three ways, (1) by unsubstantiated "gut feeling," (2) by *intra*-professional judgement based on a range of assessment information, and (3) by *inter*-professional judgement through collegial discussion (Poskitt and Mitchell, 2012).

By contrast, comparative judgements have been found to be a more reliable means of holistic assessment. Based on the insight that all judgements of quality involve comparative, tacit or explicit evaluation of assessment artefacts (Laming, 2004), comparative judgement approaches such as pair-wise comparison (Heldsinger and Humphry, 2010; McMahon and Jones, 2015) and adaptive comparative judgement (Pollitt, 2012; Bartholomew and Yoshikawa, 2018; Baniya et al., 2019; van Daal et al., 2019) have shown high levels of reliability, even when compared with assessment against pre-set criteria (Bartholomew and Yoshikawa, 2018).

Underpinning holistic or global assessment judgements are tacit, "in the head," models of quality which teachers bring to the assessment event. These "prototypes" (Rosch (1978) or "implicit constructs" (Rea-Dickins, 2004) are internal conceptions of quality as a generalised attribute, which are mobilised as standards of comparison in the course of engagement with student assessment artefacts. These internal, construct-referenced standards have been found at work in evaluative processes during the formation of teachers' assessment grading decisions (Crisp, 2010). Here, Crisp found that the "Cartesian gestalt model" (Cresswell, 1997) where an assessor "identifies the presence or absence of certain features and then combines these cues *via* a flexible process to reach a judgement of grade-worthiness" (Crisp, p. 34) best describes this judgement process of "comparing to prototypes." In this context, mental portraits of students (Yin, 2010) may also be seen as a kind of prototype in which stored impressions about particular types of

students act as a reference point for comparative judgements about students' relative strengths and weaknesses.

As described by Sadler, the formation of final overall assessment judgements is the product of reflexive interaction between global and analytical assessment:

> Experienced assessors routinely alternate between the two approaches in order to produce what they consider to be the most valid grade. . ..In doing this they tend to focus on the overall quality of the work, rather than on its separate qualities. Among other things, these assessors switch focus between global and specific characteristics, just as the eye switches effortlessly between foreground (which is more localised and criterion bound) and background (which is more holistic and open (Sadler, 2009, p. 57).

Similar two-way interactions involving descriptor interpretation, judgement negotiation, comparing across samples, differential attention to criteria and work samples, and implicit weighting criteria have been reported in detailed studies of teacher assessment decision-making (Klenowski et al., 2007; Wyatt-Smith et al., 2010).

A final consideration is a generalised model of how judgement-centred assessment operates in classroom situations. Wyatt-Smith and Adie (2021) draw on Sadler's criteria classification of explicit, latent, and meta-criteria (Sadler, 1985, 2009; Wyatt-Smith and Klenowski, 2013) to provide a realistic cyclical account of how these criteria interact during teachers' appraisal processes. In this cyclic appraisal model, teacher analytical feature-by-feature assessment arising from stated criteria interacts with reflection on a global appraisal (emergent, latent criteria) that synthesise as an overall assessment judgement according to certain meta-criteria—the knowledge of how explicit and latent criteria can be combined. Latent criteria might include global impressions such as prototypical models of quality, student mental portraits, and teachers' prior judgements carried forward over time. This process highlights the key role reflexive decision-making processes play in effective teaching and assessment (e.g., Clark and Peterson, 1984; Wilen et al., 2004; Good and Lavigne, 2017).

The dynamics of judgement appraisals and its centrality to teacher-based assessment has been well documented in studies on situated judgement practices in assessment moderation contexts (e.g., Klenowski et al., 2007; Wyatt-Smith et al., 2010; Adie et al., 2012; Wyatt-Smith and Klenowski, 2013, Wyatt-Smith and Klenowski, 2014). The notion of judgement practice however, needs broadening to better reflect the professional, epistemic and evaluative agency teachers develop through recurring classroom assessment activity. Elaborating the concept of L2 assessment praxis (Michell and Davison, 2020) as *judgement praxis* aptly describes the conscious and tacit tool-mediated, judgement-based assessment knowledge practices reviewed in this section.

Gestalts and Decision-Making

Gestalt Psychology

With its holistic view of human perception and action, Gestalt Theory and its concept of Gestalt offers insights into what happens inside the cognitive "back box" of language teacher assessment decision-making. Roughly translated as "configuration" (Jäkel et al., 2016, p. 3) or more accurately as "whole" or "form" (Cervillin et al., 2014, p. 514), the concept of Gestalt was first introduced to psychology in the late 1890s by a German psychologist Christian von Ehrenfels (Wagemans et al., 2012a,b). The concept was later extended as Gestalt Theory by Wertheimer (1912), who, together with Kurt Koffka and Wolfgang Kohler, founded the Berlin School of Gestalt psychology. These Gestalt psychologists investigated the psychology of visual perception with a view to understanding human mind and thought in its totality.

Koffka (1935, 2013) theorised the key Gestalt principles of perception organisation, namely, *similarity*—similar items tend to be viewed as a group; *prägnanz* (simplicity)—objects are viewed as simply as possible; *proximity*—items near each other tend to be categorised as a single group; *continuity*—perception favours alternatives that allow contours to continue with minimal changes in direction; *the law of closure*—the tendency of human brain to complete shapes by filling gaps in missing parts; and the *law of common fate*—"the tendency for elements that move together to be perceived as a unitary entity" (Wertheimer, 1923 as cited in Wagemans et al., 2012a, p. 1,181).

The primary principle behind the Gestalt laws of perception organisation is that the whole is other than the sum of its parts, meaning the whole should be viewed as the interwoven and meaningful relationship between parts, not simply as an addition of parts to make the whole (Koffka, 1922, 2013). Gestalt is "a whole by itself, not founded on any more elementary objects . . . and arose through dynamic physical processes in the brain" (Wagemans et al., 2012a, p. 1,175). Thus, the meaning and the behaviour of the whole is not determined by the behaviour of its parts. Rather, the intrinsic nature of the whole determines the parts (Wertheimer, 1938, 2012). This is theorised in modern Gestalt psychology as the primacy of holistic properties which cannot be perceived as individual constituents, but only by their interrelations. This means that holistic configurations dominate constituents during information processing; perceptions are constructed "top down" rather than "bottom up." In sum, the central idea of Gestalt psychology from both traditional and modern perspectives is the dominance of the whole over its parts in perceptual processing.

Gestalt in Language Teacher Cognition

Gestalt theory therefore offers valuable insights into the holistic, impressionistic aspects of teachers' language assessment decision-making. Gestalts may be understood as part of the sense-making (Kubanyiova and Feryok, 2015) or imagistic orienting activity (Feryok and Pryde, 2012) processes of language teacher cognition and can be equated with "situational representations" (Clarà, 2014) that develop through experience of the immediate demands of teaching activity to become the stock and store of teacher knowledge practice.

Gestalts play a key role in Korthagen's model of teacher learning as situated cognitions:

> [A Gestalt is] a dynamic and constantly changing entity, [that] encompasses the whole of a teacher's perception of the here-and-now situation, i.e., both his or her sensory perception of the environment as well as the images, thought and feelings,

> needs values, and behavioural tendencies elicited by the situation (Korthagen, 2010, p. 101).

The process of Gestalt formation is the result of a multitude of everyday encounters with similar types of classrooms situations. Korthagen's three level model of Gestalt formation from concrete experiences to schematisation to theory formation and then subsequent reduction of schema and theory elements as higher-order Gestalts highlight teaching as a Gestalt-driven activity in which Gestalts are triggered by certain classroom situations when sufficiently rich schema has been developed. In this way, Gestalts are both a key resource and driver of teacher cognition, learning and expertise available for recognition and recall to guide classroom decision-making (Klein, 1997).

Gestalt Cognition in Clinical Judgement

Teacher assessment judgement is akin to clinical judgement in the medical professions, specifically in the areas of diagnosis, therapy, communication and decision-making (Kienle and Kiene, 2011). As with teacher assessment judgements, doctors apply their connoisseurship, expertise and skills to establish

> a relationship between the singular (everything the evaluator knows about a particular individual) and the general (formal and tacit professional knowledge, as well as institutional norms and rules) in order to formulate the most appropriate course of action possible (Allal, 2013, p. 22).

Gestalt cognition lies at the heart of clinical judgement. Often manifesting as a "hunch," it enables expert practitioners to swiftly interpret situations, develop a global impression of a patient's health status, make causality-effect judgements and decide on appropriate treatments. Gestalt-based predictive causality assessments develop over time through repeated practice, experience, knowledge and critical reflection:

> Personal experience can translate into Gestalt cognition, which can be recast into the logic of tacit thought, and can eventually translate into the tacit power of scientific or artistic genius (Cervillin et al., 2014, p. 513).

Recently, there has been something of a reassessment of the value of Gestalts in clinical decision-making. The application of "evidence-based" scientific methods for evaluating clinical reasoning has not necessarily lead to better health outcomes and, unlike clinical judgement, "gold standard" cohort-based, statistics-driven, probabilistic research such as randomised controlled trials cannot determine effective treatment outcomes for *individual* patients (Kienle and Kiene, 2011). Gestalt cognition has been shown to enhance the effectiveness of medical practices such as physical examination, electrocardiogram analysis, imaging interpretation and difficult patient diagnoses (Cervillin et al., 2014), and, in the pandemic context, Gestalt-based clinical judgements in virtual, online consultations (Prasad, 2021).

Gestalt as Heuristic Insight

Extending Gestalt theory, Laukkonen et al. (2018, 2021) have highlighted "insight" at the heart of Gestalt cognition by drawing attention to the insight experience associated with eureka (aha!) moments and its effects on the cognitive-emotional appraisal of ideas and decision-making. Phenomenologically, these "feelings of insight" are often experienced as a sudden illumination after an extended incubation period of problem solving. Often associated with inherent confidence (Danek and Salvi, 2020), these powerful feelings "act as a heuristic signal about the quality or importance of an idea to the individual" and "play an adaptive role aiding the efficient selection of ideas appearing in awareness by signalling which ideas can be trusted, given what one knows"(p. 27).

The phrase "given what one knows," is a major caveat since "false eurekas" can be elicited experimentally and false insights occur when an idea is consistent with one's knowledge but inconsistent with the facts. If one's implicit knowledge structures are invalid, then insights arising therein will also be invalid. Such Gestalts then are no guarantee of truth but are only as solid as the knowledge and expertise that lies behind it. The implication for language assessment decision-making is clear—to be established as trustworthy, such insights need to be followed by, and subject to, reflection, analysis and verification.

MATERIALS AND METHODS

Research Design

This study was part of larger mixed-method study (Johnson and Christensen, 2010) on variability in teachers' oral English language assessment decision-making. The study aimed to provide insight into this process through the following research questions.

1. What are the processes of teacher decision-making when assessing student's oral language performances?
2. How trustworthy are teachers' assessment judgements?

The study was conducted in three stages: (1) a participant project information, consent and assessment training session in which a *questionnaire* was used to collect background information from the participating teachers, (2) a *teacher assessment activity* in which teachers watched a set of videos of students' performances and assigned scores to student performances and (3) a *retrospective think-aloud activity* and follow-up *semi-structured interviews* to further investigate explanations of teachers' decisions and justifications.

The design of this qualitative study of teachers' assessment decision-making reflects Vygotsky's process analysis which recognises that, as "any psychological process. . . a process of undergoing changes right before one's eyes" (Vygotsky and Cole, 1978, p. 61). Consequently, "the basic task of research. . .becomes a reconstruction of each stage in the development of the process"(p. 62) "in all its phases and changes—from birth to death . . . to discover its nature, its essence, for it is only in movement that a body shows what it is" (p. 65).

Participants

Participants were selected using convenience sampling. Currently practicing EAL/D teachers from the state professional association were invited to take part in the study. Twelve teachers took part in the full research study. Teachers were drawn from primary

TABLE 1 | Participants' background information.

Teacher	Age	Current teaching position	TESOL qualifications	Teaching experience (years)	Languages of students taught
A	56+	Consultant	Yes	16+	Chinese
B	41–55	Secondary	Yes	11–15	Chinese, Korean, Vietnamese
C	26–40	Secondary	Yes	16+	Chinese
D	26–40	Primary	Yes	11–15	Thai, Chinese, Arabic
E	41–55	Secondary	Yes	16+	Vietnamese, Arabic
F	26–40	Consultant	Yes	6–10	English
G	56+	Consultant	Yes	16+	English
H	56+	Consultant	Yes	16+	English
I	56+	Primary	Yes	16+	Chinese, Arabic, Persian
J	41–55	Primary	Yes	11–15	Chinese, Spanish
K	41–55	Secondary	In progress	11–15	LBOTE
L	56+	Primary	Yes	11–15	Hindi

and secondary levels in NSW: seven from the government school sector, two from the Catholic school sector and one from the independent school sector. Background information about participants was collected from a questionnaire which was also used to obtain teachers' consent to participate in the training workshop and the assessment activity. The results of background information questionnaire are shown in **Table 1**.

As shown in the table, all participants were highly experienced EAL/D teachers, with half teaching for more than 16 years, five teaching for between 11 and15 years and one teaching for between 6 and 10 years. Four participants were EAL/D consultants, who worked closely with EAL/D teachers and learners at both primary and secondary levels. Teachers had experience in teaching students from diverse language backgrounds. With one exception, all participants had TESOL qualifications in addition to their general teaching qualification. All participants were female.

Teacher-Based Assessment Activity

A teacher-based assessment activity was conducted immediately after the questionnaire administration (**Table 2**). The activity replicated the TEAL Project professional learning workshop design and, as the teachers did not know the students presented in the video stimulus, assessment took place "Out of Context" (Castleton et al., 2003; Wyatt-Smith et al., 2003).

Participants were asked to view three video samples of student assessment task activity and score their oral language performance against task-specific assessment rubrics. Descriptions of video samples are presented in **Supplementary Appendix A**. The rubric comprised an equally weighted, four-level rating scale with each level indicated by a set of criteria across four different linguistic categories—communication, cultural conventions, linguistics structures and features and strategies (**Supplementary Appendix B**).

After a practice run, teachers were asked to highlight the performance descriptors that matched the performance they observed in silence; then decide on students' performance levels in a scale from 1 to 4. In addition, they could add any comments they thought would justify and support their final decisions they made against the student. Teachers were then shown the video of each student sample twice. During the first time watching the first student sample, teachers were encouraged not to refer to the criteria; however, they could use the criteria sheet the second time. Teachers' task assessment scores are recorded in **Supplementary Appendix C**.

Immediately after finishing scoring for each student performance sample, teachers were asked to compare and discuss their assessment results in groups of three before they moved on to another task. Discussion focused on the two guiding questions: "Compare your responses. What was similar and what was different? Why did you have differences?"

In the next stage, after teacher assessments were examined for variability, teachers were individually followed up and were asked to justify their assessment decisions. Immediately after their oral justifications, teachers were interviewed with a view to obtaining more insight into their decision-making process.

Materials

Three tasks were selected from a bank of twenty one oral assessment tasks developed for the TEAL assessment project in Victoria accessed from the project website at http://teal.global2.vic.edu.au/. These tasks were designed to assess upper primary and secondary students' English language performances, meaning that both primary teachers and secondary teachers can suitably use these tasks to evaluate their student outputs. Detailed descriptions of the video stimulus material are summarised in **Table 3**.

Data Collection

Data collection was conducted *via* a 3-h accredited professional development workshop delivered and trained by an assessment specialist. Teachers signed up for either a morning session or an afternoon session. Methods employed for data collection are outlined in the previous research design section.

Think-Aloud Protocols and Interviews

Think-aloud methods have been widely employed in studies in language assessment (e.g., Cumming, 1990, 2002; Weigle, 1998; Barkaoui, 2007). *Retrospective* think-aloud protocols rather than *concurrent* think-aloud protocols were used as the latter poses a

TABLE 2 | Stages of data collection.

Stage	Description	Materials	Data collection
Participant assessment training session	Research project information and consent. Introduction to TEAL resource, practice assessment with assessment tools	TEAL videos, assessment task rubrics, scoring sheets	Participant questionnaire
Participant assessment activity	Participants view student video performance twice and individually rate each student against task assessment rubric, then compare their decisions after each task in groups of three	TEAL videos, assessment task rubrics, scoring sheets	Audiotaping and transcription
Retrospective think- aloud activity and follow up semi-structured interview	Participants review the videos and their score sheets, then justify their ratings of students' performances	TEAL videos, assessment task rubrics, scoring sheets, interview question guide	Audiotaping and transcription

TABLE 3 | Video oral language work sample material.

Assessment task	Description	What shown on the video	Language being assessed	Students being assessed
Task 13: Choosing a gift for a character	task requires students to discuss characters and events in a familiar literary work to reach agreement about a suitable gift for a character in the story	students participating in collaborative discussion with peers	listening and responding, interacting and negotiating	a Year 10 female from China
Task 19: A book or film review	task requires students to describe plot, characters, themes and issues and provide evaluative comments and a personal response to a novel or a movie in response to questions from a classmate or teacher	shows two students giving a brief spoken report and personal response	oral presentation of information	a Year 8 male from China
Task 21: Job interview role play	task requires students to participate in an interview about themselves in relation to a hypothetical job	shows a student answering questions from an adult male interviewer and talking about themselves in a positive, culturally appropriate way	listening and responding, interacting and negotiating, oral fluency and flexibility	a Year 8 male from Mongolia

complex and difficult multitasking challenge for teachers while the former has been reported to increase teachers' verbalisation by reducing their cognitive load through spacing viewing and scoring activity from explaining assessment decisions (Bowers and Snyder, 1990; Van Den Haak et al., 2003).

Teachers were individually invited to complete retrospective think-aloud protocols which were implemented 1 week after the teacher-based assessment activity. During the think-aloud protocol, teachers viewed their scored criteria sheets and watched the videos of student speaking tasks again, and explained what they had thought and decided in the teacher-based assessment activity. After completing their think-aloud protocols, individual teachers were followed up in semi-structured interviews in order to obtain rich data about their assessment justifications and decision-making. An interview guide consisting of predetermined structured questions and follow-up open-ended questions was used (**Supplementary Appendix D**). The interview questions were divided into three major categories to cover information about teachers' assessment confidence, processes and biases. Qualitative interviews were chosen for their value in eliciting in-depth information about social processes, and the "how" of psychological phenomena. All teacher discussion and interviews were audiotaped with the consent of the participants and later transcribed.

Data Analysis

Data from the three data sets below were analysed and triangulated with a view to identifying interaction between holistic and analytical assessment processes, the role of Gestalt-like judgements in these processes, and patterns in teachers' assessment decision-making and their relative trustworthiness.

Analysis of Questionnaire Data

Background information collected from 12 participants through questionnaire. Responses from close-ended questions were turned into numerical data and analysed using descriptive statistics methods through the statistical computer software SPSS. The questionnaire data were then analysed in conjunction with the assessment data. Findings from these analyses were triangulated with the information obtained from the think-aloud protocols to answer the second research question.

Analysis of Teacher Assessment Scores

To analyse teacher variability and consistency, mean score calculations were conducted on teacher grade scores. Each teacher marked three student outputs using the criteria including seven assessment categories. Individual marks were taken as separate subsamples for data analysis. Teachers' individual judgment scores in each category were therefore considered as a

distinct variable with each teacher assigned 21 scores, making up 252 observations. This number of observations was large enough for the purpose of analysis. However, given this was still a fairly simple data set, all data collected from the assessment activity were manually calculated. For the purpose of calculation, data were first modified prior to primary analyses being conducted.

Analysis of Group Discussion and Interview Data

Transcriptions of the post-assessment group discussions were analysed to design the interview questions for the follow-up interviews to the retrospective think-aloud activity. Key themes and subthemes from all three sources were iteratively identified and triangulated (Miles and Huberman, 1994; Esterberg, 2002; Nunes et al., 2019). The coding scheme suggested by Cumming et al. (2002) was adopted to identify influential themes, with data coded both according to predetermined themes identified in the literature and using grounded theory (Glaser and Strauss, 1967), used mainly to untangle issues of outlier assessment behaviours. To facilitate coding and coding management, a computer program NVivo version 10 was used. Aptly for this study, researcher immersion in the data led to a gestalt of the assessment pathway model which the researcher subsequently analysed, verified and refined against the data.

RESULTS

Qualitative Analysis of Teacher Assessment Pathways

Analysis of teacher discussion, think-aloud and interview data identified the key role of teachers' first impressions, or Gestalts, in assessing students' oral performances. These Gestalts were found to determine the nature and trustworthiness of teachers' final assessment judgements through one of three identifiable assessment decision-making pathways—balanced, conflicted and unbalanced. These Gestalt-based assessment pathways were further tested against the data and refined as the model of Gestalt-based language assessment decision-making shown in **Figure 1**.

This section presents the analysis of the verbal data from teacher discussion, think-aloud activity and interviews in each of the three assessment pathways of the model in order to show how teacher's assessment decision-making unfolds in these pathways. From the twelve participants, three groups were identified in relation to each of the assessment pathways. Six teachers were found to have formed trustworthy, balanced assessment judgements through a strong Gestalt/high reflexivity pathway; one teacher formed unconfident conflicted assessment judgements through a weak Gestalt/low reflexivity pathway; while five teachers formed suboptimal trustworthy, unbalanced assessment judgements through a strong Gestalt/low reflexivity pathway.

Balanced Assessment Pathway

The balanced assessment pathway was identified as a highly reflexive assessment decision-making process in which teachers arrived at a trustworthy, "on balance" judgement of students' language skills as a result of robust interrogation of their strong initial assessment Gestalt and the adequacy of related available assessment information. This assessment pathway unfolded in three stages—formation of a strong initial assessment gestalt, robust self-interrogation and a final balanced assessment judgement. Teachers C, E, F, G, K, and L were in this pathway group.

Stage 1: Formation of a *strong initial assessment Gestalt*

After watching the videos, teachers in the balanced assessment pathway group reported that they formed clear impressions of the relative strengths and weaknesses in the talk of the three students being assessed. Certain features of each of the students' talk stood out and gave them an immediate and generalised sense of where students might be placed on the task assessment performance levels. Teacher's first impressions were thus triggered and formed by students' individual and comparative language performances and continued to influence subsequent assessment decision-making.

For example, **Teacher C** reported her initial impression of Student 1 was that "her oral language was clunky and . . . forced." She was impressed nevertheless, with the student's understanding of the content, noting: "she developed really good ideas." Her clear impressions of Student 2 were formed in the context of comparison with Student 1:

> He had a really sophisticated sort of grasp of informal English. You know, he spoke confidently, he was using it really well, he wasn't looking . . . whereas, yeah, the girl was really clunky, as opposed to [Student 2].

As with other teachers in this group, Teacher C found that Student 2's communication and interpersonal skills had a positive impact on her. Like the other teachers, her first impression about the third student was an overwhelmingly positive one of oral fluency:

> He's mastered the pronunciation, the American pronunciation really well. So, if I saw him I'd go yeah, automatically, he's fine for entry, his oral language is fine.

Teacher E's first impressions went beyond Student 1's apparent disfluency. She was impressed by the way the student took part in the conversation (e.g., starting and maintaining the conversation):

> Because it's also easy to be distracted by the negatives, but the detail I think, and her case, she didn't leap into it. You could see that as a negative but actually I could see that she was just thinking things through carefully before she spoke.

Student 2's communication and interpersonal skills also impressed her and elicited a high score. She rated Student 3 as quite a competent speaker:

> He is obviously quite articulate and his grammatical features I thought were quite good and his text structure is quite high. I thought he would come out on top.

By contrast, when talking about her first impressions, **Teacher K**, focused on what she thought were salient aspects of Student 1's personality:

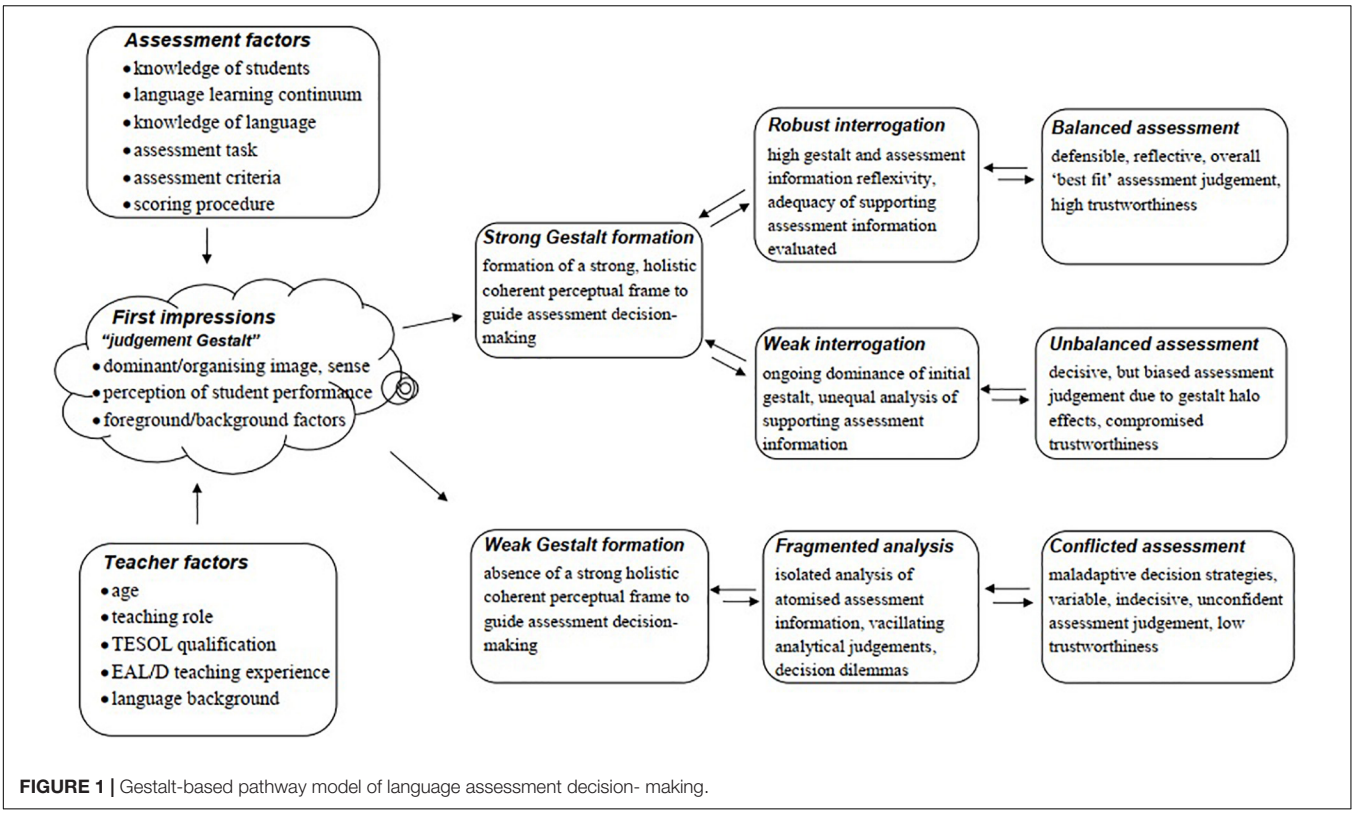


FIGURE 1 | Gestalt-based pathway model of language assessment decision- making.

> I just liked her assertiveness but that could be . . . because I just appreciated the fact that even though she is lacking a little bit of, I guess fluency with her spoken text, she really put herself out there and she butted in a bit. I liked that.
>
> She got the conversation going. It would have come to proactivity. So, I thought she was very proactive.

She also indicated that she was impressed with Student 2's pragmatic approach which she thought "was a major strength for him," adding that she felt he was very good at engaging his counterpart and eliciting details from his partner. Like the other teachers, she was impressed by his communication and conversation skills compared with Student 1:

> He did it in a friendly way. He didn't do it in the same way as the first student. He was really good at keeping the conversation going, the dialogue going.

From these representative accounts, it is evident that the teachers quickly formed strong holistic appraisals of students' speaking skills from their comparative viewing of students' oral language performances. Triggered by observed language features that teachers considered to be salient, these judgement Gestalts arose as unified configurations of inseparable elements of language features, assessment task performances, student intentionality and agency and inferred or imputed personality characteristics. Formed without reference to any pre-specified criteria, they were frequently described by the teachers in terms of their immediate impact, most commonly as "being impressed."

Stage 2: Robust interrogation of initial assessment Gestalt and supporting evidence

In this stage, teachers interrogated their initial judgement Gestalts of student performances as well as the information gained from analysing the task assessment rubric or from further reflection on student observed language performance. This stage marked the shift from, and interaction between, the "fast" thinking of Gestalt appraisals and the "slow" thinking of rubric analysis.

The additional time needed for analytical consideration of assessment rubrics is prominent in **Teacher C's** response. She commented that the student dialogue gave her time to read through and reflect on the criteria while the non-assessed students were speaking, "The fact that it was dialogue was quite good because it forced you to also reflect back on what you had ticked and things like that." She added if the dialogue was shorter, "1 min or even 30 s," she may have been forced to make an inaccurate assessment.

This move to analytical thinking around the task assessment rubrics allowed interrogation of and reflection on teacher's initial assessment Gestalts. **Teacher E** thought she could not rely solely on her positive first impressions of the two students to judge how well they were performing their task but would "have to stick on the indicators" in the assessment rubric. **Teacher F** expressed a similar view, reiterating that even though her initial impressions suggested the students were positive, she could not provide an accurate score without using appropriate assessment criteria:

> . . . first impressions may be good, because I may have . . . unconsciously I have criteria set in my mind. I go, "oh that's good." But when I look at the assessment criteria, the assessing criteria I know I need to follow this standard.

A key outcome of this interaction between Gestalt- and rubric-mediated judgements was the verification of the teachers' initial judgement Gestalts and a confident grading decision. **Teacher L** reported:

> But then, when I see the criteria, you know, this specifies where they are at. Because when I look at them, it's just general. I can't find where do I need to assess them. And then when I see this, "oh this is where they should go."

Similarly**, Teacher K's** analysis of the assessment criteria confirmed her first impressions of Student 3 "as more fluent and more experience[d] in the use of English."

> I still relied upon the criteria. I was really pleased when I started doing it that it was quite accurate in its format. That it came up, based on my note-taking it came up at a higher level than the other students. I was quite thrilled about that and I thought this is actually quite a helpful tool.

From these accounts, it is clear that the teachers placed great value on using assessment rubrics as aides to reflect on their initial assessment Gestalts and ensure accurate and reliable language assessment. The stage highlights the extra time needed for "slow" analytical engagement with assessment criteria in contrast to the "fast" immediacy of judgement Gestalts. Teachers' initial Gestalts did not disappear, however, but remained as coherent organising frames guiding assessment decision-making.

Stage 3: Further reflection on supporting evidence and balanced assessment

In this final stage, teachers form an "on balance" assessment judgement from the interaction between, and interrogations of, their assessment Gestalt and rubric-mediated assessment information. This stage is characterised by high reflexivity motivating teachers to interrogate the relevance and adequacy of existing information and seek out additional "missing" evidence that enables them to form a sound, confident overall judgement of students' language skills.

Teacher C's comments on deciding on Student 1's task performance level reflects high level awareness of the common mistakes teachers make when assessing students' oral performance. This awareness impels her, not only to interrogate available information, but also to seek out and weight further necessary evidence about student's real language abilities.

> As teachers, when you're assessing students, you've got to be mindful of how . . . because we do get fooled by students who talk the talk really confidently and things like that. Whereas the little girl [S1] her expression was not so great, but she had some really good ideas, she had some really good understanding of the text. So, I think you've got to be really careful, and if you're assessing for understanding you've got to make sure that that is weighted more and that teachers can see that.

Similarly, **Teacher E** was aware that an overall assessment judgement needed to take account of student performance at different levels across different skill areas. Despite Student 1's strategic competence, enthusiasm and engaging conversation, she required further information to form a comprehensive overall judgement of the student's oral language ability:

> It helped to inform that first communication because it was an overall judgement about the type of communication skills she had, but I don't think it affected the other aspects in terms of her strategic competence because I knew I had to look for other features.

Her reflexivity was also evident in deciding on Student 2's performance level. Although Student 2's communication and interpersonal skills impressed her and suggested a high rating, she was prepared to look beyond surface-level phenomena:

> You have to step back and listen to the content and actually he didn't have a lot of content although he did have some good vocabulary, so . . . but his grammatical features he had some grammatical inaccuracies which were easy to overlook because of his fluency.

The balanced assessment judgements achieved by the teachers in this group was an outcome of holistic and analytical assessment appraisals which were both subject to robust interrogation, including considerations of necessary supplementary evidence. This process of sustained meta-reflection made possible confident and trustworthy overall teacher assessments of students' language skills.

Conflicted Assessment Pathway

The conflicted assessment pathway was identified as a decision-making process in which the teacher was unable to make an "on balance assessment" of students' oral language skills due to a weakly formed assessment Gestalt and a resulting fragmented analysis of isolated language elements from the task assessment rubric. The conflicted nature of the assessment was evident in the teacher's "torn" vacillation between equally weighted analytic elements of the students' performance and her lack of confidence in her final assessment judgement. This assessment pathway unfolded in three stages—a weakly formed initial assessment Gestalt, fragmented analysis and a final conflicted assessment judgement. Only one teacher, **Teacher D**, was found in this assessment pathway.

Stage 1: Weak formation of initial assessment Gestalt

Like the teachers in the balanced assessment pathway, this teacher observed the relative strengths and weaknesses of students' oral language performances. Unlike these teachers, however, she did not form an overall perceptual frame that could provide a central, coherent reference point for judging students' oral language skills.

This weak Gestalt is indicated by her "split," indecisive appraisals of Student 1. On the one hand, her responses during the group conversation were "rather structured, formulaic and stilted," but, on the other hand, she "accurately uses formulaic structures to indicate turn taking." Further, the initial impressions gained from comparing the oral language skills of Students 2 and 3 were somewhat superficial and were not interrogated

> He (Student 2) was definitely better than the first one (Student 1). And a lot of that had to do with the spontaneity and colloquialisms that he had.
>
> He (Student 3) was self-correcting as well which was very good. They all did a bit. And it also helped that he's developed a bit of an accent as well that is a native like [sic] accent. It sounded quite American.

Stage 2: Fragmented analytical assessment

As with the previous teachers in this stage of the assessment pathway, Teacher D shifted her attention to the task assessment rubric. However, the conflict between her initial (superficial) impressions and assessment criteria soon became apparent:

> Like I said before, for instance, that last student, well the second student, he was just so funny, and because he's so confident . . . then the criteria grounds you.

The absence of a strong guiding assessment Gestalt led to atomised analytical assessment characterised by a criteria-by-criteria examination of the language descriptors on the task assessment rubric and rating decisions without reference to an overall appraisal:

> You start looking at, what about their verb endings, are they using modal verbs, are they just using formulaic language. I think that is very important to come down.

Similarly, students' "borderline" performances are resolved without reference to holistic appraisals, "if I had to give the students a one to four, they'd all probably be a bit higher."

When asked whether her first impressions influenced her assessment decision-making, teacher D was uncertain and non-specific, "Yes, well, quite a bit I think." Her further reflections on this issue were similarly non-specific:

> I think, as a reflective teacher, that I would have to be a bit dishonest to say that I do not have biases. And maybe they're not conscious, but I think everybody does.

In the absence of a strong overall guiding assessment Gestalt, Teacher D's assessment becomes little more than atomised "criteria compliance" (Wyatt-Smith and Klenowski, 2013) where equally weighted descriptors foster conflicted and vacillating assessment decision-making.

Stage 3: Conflicted assessment judgement

In the absence of a strong guiding assessment Gestalt, Teacher D resorts to contradictory or inconsistent decision-making strategies and final indecision, when pushed.

For example, when grading Student 1's performance, she decided that this student was halfway between a level two and three: "If I can't decide I should always assess them down." This strategy was contrary to what she had said earlier when she indicated that she would give higher scores for students on the borderline. However, in the end, she applied her own "middle halfway" decision-making strategy:

> That's how I reached that decision . . . I went "Okay, she's halfway in-between so I'll go for two.

When assessing Student 2's performance, she was torn between giving a global rating of student language competence based on her initial comparison with the previous student, and her reading of the assessment criteria. Although she felt the student was very confident and she wanted to rate him at level four, "in the end I felt that I couldn't, based on the criteria." Similarly, when deciding on Student 3's performance on one of the language skill areas, she could not arrive at a final overall assessment judgement:

> I couldn't decide . . . I gave him two and then I changed it back to a three and I couldn't really decide for that one. And that probably dragged him down a little bit as well. I think if I'd been confident that that was a level three, then maybe I could have pushed him up a bit more.

In this final stage of the conflicted assessment pathway, then, Teacher D's uncertainty and indecision fostered maladaptive decision-making strategies which undermined the confidence and trustworthiness of her final assessment judgements.

Unbalanced Assessment Pathway

The unbalanced assessment pathway was identified as an assessment decision-making process in which teachers were unable to make an "on balance assessment" judgement due to inadequate interrogation of a strong initial assessment Gestalt. This pathway resulted in decisive but unbalanced assessment judgements with sub-optimal trustworthiness due to halo effects associated with the persistent strength of the initial Gestalt. This assessment pathway unfolded in three stages—formation of a strong initial assessment Gestalt, weak self-interrogation and a final unbalanced assessment judgement. Five participants, Teachers A, B, H, I, and J were in this pathway group.

Stage 1: Formation of a strong initial assessment Gestalt

As with the previous two decision-making pathways, teachers' first impressions of students' oral communication skills were formed from viewings of their task performances. In this pathway, teachers' initial assessment Gestalts were associated with perceived aspects of students' personalities related to their language performance:

> At first she was very confident. She presented a very diligent student who'd really gone over the material. She's obviously familiar with that. Her articulation, you know she opened her mouth and articulated (Teacher A about Student 1).
>
> With the girl, I was impressed at how she did throw a bit of insight into the ideas of the film. It wasn't just a black and white . . . she was able to counteract. I thought that was really good. She was clever, I thought (Teacher B about Student 1).
>
> [he] was a very skilled communicator. and very engaging and, you know, he's got a lot of personality, very interested in people. He was very observant, he's watching the person he's communicating with and reading memos (Teacher A about Student 2).

Task salient aspects of students' personalities are foregrounded and teachers' attention is drawn to the way students take charge of, lead or sustain the group conversation:

> When you look at the first group, the three students sitting there together, one thing I did like [was] how the girl held the conversation . . . So, I think that would influence me in terms—even though I know we're probably meant to assess language skills, but I think she was very good, and that's why I would be more influenced for her (Teacher B).

> She clearly knows how to interact in a discussion. So, her strengths are that she knows what an oral discussion is all about (Teacher H about Student 1).
>
> She was the type of student who would take a leadership role in any group work (Teacher I about Student 1).
>
> He's confident. He appropriately avoids negotiating and communicating. I think it's quite clever. I'd do the same thing. I have a very sustained conversation (Teacher H about Student 3).

Teacher I believed that she might score Student 1 higher because "she seemed to take charge and seemed to be very competent." Teacher H found that Student 2 had "an engaging personality in an oral discussion" and that what this student really needed was vocabulary to be "a very articulate, engaging speaker."

Conversely, Teacher B's first impression of Student 3's job interview performance was affected by the student's lack of interaction and engagement, "he was a little boring in his responses." Consequently, she focused on his drawbacks such as "his pronunciation of words by default." As seen in the other assessment pathways, these Gestalts were stimulated by a comparative assessment of students' strengths and weaknesses:

> One of his strengths was in the way he spoke. He did sound colloquial, but because it wasn't too formal, and I think that's how your attention [was] a bit with his conversation, [not] with the girl (Teacher B about Students 2 and 1).
>
> The girl had good answers. She knew what she was talking about. She had a lot of knowledge about the characters. More so than what he had . . . but he displayed more confidence in the way he was speaking than the girl. She sat quite still, whereas he was leaning all over, which I think is a street, smart kind of kid. He didn't have the formality in the same way as the girl did, but that could be part of his personality as well, because people have different kinds of personalities (Teacher B).
>
> I know you're not meant to compare students. You're not meant to compare, but it is really hard not to (Teacher B).

While the origin, formation and nature of teachers' assessment Gestalts parallel those in the balanced assessment pathway, what is noticeable in this pathway is their relative strength and power associated with perceived student personality traits. This strength persists throughout the next two stages and overwhelms and sidelines any robust interrogation required to form balanced assessment judgements.

Stages 2 and 3: Gestalt dominance, weak interrogation and unbalanced assessment judgement

These stages are characterised by the continuing dominance of teachers' initial Gestalts with weak, unequal interrogation of those Gestalts and related assessment rubric information. Teachers' first impressions of students' performances remain unchanged and persist as the dominant influence on their final assessment judgements. This Gestalt dominance is particularly evident in teachers' recognition of the continuing influence of their first impressions on their assessment thinking.

Gestalt dominance can be seen in the persistence of **Teacher B's** first impressions of Student 1 and their acknowledged influence on her final assessment decision even after considering other students' performances:

> When you look at the first group, the three students sitting there together, one thing I did like [was] how the girl held the conversation . . . So, I think that would influence me in terms—even though I know we're probably meant to assess language skills, but I think she was very good, and that's why I would be more influenced for her.

Teacher I's account highlights how holistic assessment judgements ultimately override or sideline analytical ones during grading decisions. After viewing Student 1's performance a second time, the teacher noticed that she had not realised or had ignored grammatical issues in his performance on the day "because she was providing so much information and doing it reasonably articulately." Nevertheless, her overwhelming impression that Student 1 "seemed to take charge and seemed to be very confident" in the conversation dominated and led her to believe that she might have given the student a higher score.

Similarly, her initial positive impressions of Student 2's performance persisted unchanged, despite noticing his limited talk time and several speech problems:

> He had a whole lot of the non-verbal[s] and his . . . he was the perfect talk show host. . . . and he had a lot of the . . . even the gestures and the . . . and the demeanour of a talk show host in talking into an interview . . . into an interview guest.

Remaining front-of-mind, the student's overall communication and conversation ability "would have influenced me, then." On further analysis, she identified several weak points in the student's talk but did not mark him down, but instead gave him "a relatively high score," weakly justifying, "I might have been feeling very generous that afternoon."

In this assessment pathway, then, teachers' first impressions about students' oral language performances play *the* decisive role in forming their final assessment judgements. These assessment judgements were unbalanced because teachers' reflexivity was not adequate or equal to the task of interrogating a dominant assessment Gestalt or related assessment evidence. As a result, trustworthiness of final assessment judgements is compromised by "halo effect" biases chiefly associated with student personality factors.

Quantitative Analysis of Teacher Assessment Variability and Consistency

Quantitative analysis of teacher assessment variability and consistency was undertaken to complement and check the qualitative findings of the study. The relative trustworthiness suggested by each of the teacher assessment pathways was specifically investigated through quantitative analysis of teacher assessment variability and consistency. Here, trustworthy assessment processes are identified as those that produce consistent results, when administered in similar circumstances, at different times and by different raters. It was found that quantitative analysis for both teacher assessment variability and consistency confirmed the relative trustworthiness of each of the teacher assessment pathways suggested in the qualitative analysis.

TABLE 4 | Teacher assessment variability and consistency by decision-making pathway.

Assessment outcome	Teachers	Variability			Consistency		
		S1	S2	S3	S1	S2	S3
Balanced	C	4.0	3.0	3.0	0.55	0.68	0.58
	E	2.0	3.0	3.5	0.76	0.87	0.39
	F	2.5	2.5	3.5	0.43	0.27	0.42
	G	2.5	3.0	3.5	0.57	0.49	0.51
	K	2.0	2.0	3.0	0.67	0.70	0.51
	L	3.0	2.0	3.0	0.74	0.75	0.56
	Mean	**2.5**	**2.6**	**3.25**	**0.62**	**0.63**	**0.50**
Conflicted	D	2.0	3.0	3.5	0.64	0.42	0.54
Unbalanced	A	4.0	3.0	3.5	1.10	0.63	0.32
	B	4.0	3.0	3.0	1.12	0.70	0.46
	H	2.5	3.0	3.5	0.43	0.27	0.42
	I	4.0	2.5	3.5	0.69	0.51	0.39
	J	2.0	3.0	3.5	0.64	0.47	0.32
	Mean	**3.3**	**2.9**	**3.4**	**0.80**	**0.63**	**0.38**
	Overall mean	**2.80**	**2.71**	**3.42**	**0.69**	**0.61**	**0.45**

Variability

Assessment variability is measured by the degree of difference between the mean score and the observed score and the mean scores are different for each student. This means that the variable behaviour of that teacher was stable at different times, tasks and students and, thus, predictable.

Table 4 shows the variations in comparative means of actual scores assigned by each teacher for the performance of each of the three students according to balanced, conflicted and unbalanced assessment outcomes.

In relation to assessment of **Student 1's** performance, teachers who produced unbalanced assessment judgements were found to give this student the lowest scores. The mean of actual scores by this group was 3.3, compared to the overall variability mean of 2.8. On the other hand, teachers who produced balanced assessment judgements assigned the highest scores to this student with a mean score of 2.5. The teacher producing conflicted assessment judgements tended to show most variation in her score for this student. Her assessment was significantly lower than the overall mean score at 2.0, indicating she gave the lowest score to this student.

In relation to assessment of **Student 2's** performance, teachers with balanced assessment judgements showed the least variation overall and gave more reliable scores than those in the other two groups, with the mean score at 2.6 compared to the overall mean score of 2.71. The conflicted assessment judgement teacher gave the lowest score at 3.5, meaning that her assessment for this student showed the widest variations. Assessments by teachers with unbalanced assessment judgements were a fraction higher than the overall mean score, 2.9 compared to 2.71, indicating that their assessment of this student was slightly harsh.

In relation to **Student 3**, teachers in unbalanced assessment group were found to give the most reliable score. Their mean score of 3.4 against the overall mean score of 3.42 indicated that their assessment had the least variation. Giving a slightly higher score than the overall mean score, 3.5 compared to 3.42, the teacher with conflicted assessment judgement was slightly more generous than the other assessor groups. Conversely, the mean score of teachers with balanced assessment judgements was the lowest at 3.25, meaning that their scoring for this student was comparatively stricter.

To sum up, in relation to assessment variability for individual student performances, teachers from the balanced assessment group were generally more reliable language assessors than those from the conflicted and unbalanced assessment groups. Further, certain patterns in assessment rigor were identified from the cross-student assessments of teachers in the conflicted and unbalanced assessment groups. While the conflicted assessment teacher tended to be increasingly generous in her assessments, the unbalanced assessment group's assessments fluctuated across students but always remained above the overall mean score.

Consistency

While variability indicates whether teacher assessments are "hard" or "soft," consistency describes the degree of agreement i.e., accurate and stable assessment, that a teacher achieves over different times or in different conditions (Luoma, 2004; Taylor, 2006). Ideally, it is expected that teachers score student performances in the same way. A student should receive a consistent score no matter how many teachers are involved in assessing their performance. By receiving consistent scoring from different teachers, students' ability in a task is fairly reflected and the result can be relied on for fulfilling the purpose of the assessment task.

Consistency is measured by the degree of difference between the mean score and the actual scores assigned by teachers—the smaller the difference, the more reliable the assessment. Consistency *for individual students* is indicated by the extent

to which an observed score given by a teacher to a student is close to the mean score. Consistency *across students* is indicated by the extent to which an observed score by a teacher is close to the mean score consistently across students. **Table 4** shows differences in assessment consistency between teachers in the three assessment pathway groups.

In relation to **S1's** performance, teachers producing unbalanced assessment judgements tended to have the least consistent assessments, followed by the conflicted assessment teacher and teachers in the balanced assessment group. For example, the difference between the unbalanced assessment group's average assigned score for Student 1's performance and the overall mean score by all 12 teachers was 0.80, followed by 0.64 and 0.62 for the conflicted and balanced assessment teachers, respectively. Thus, teachers producing balanced assessment judgements assigned the most consistent scores for this student's oral output.

For **S2's** performance, the conflicted assessment teacher produced the most consistent assessment with a difference of 0.42 between her score and the mean score. Teachers producing conflicted and unbalanced assessment judgements showed the same degree of consistency in their assessments of S2's performance, namely 0.63.

A different situation was observed among the three groups regarding consistency in assessing **S3's** oral output. Here, the unbalanced assessment teachers were found to make the most consistent assessments at 0.38, while those from balanced and conflicted assessment groups followed at 0.50 and 0.54, respectively. Overall, the unbalanced and balanced assessment teachers were the most consistent in their cross-student assessments, with the conflicted assessment teacher with the least consistent assessment.

It is also worth examining the internal consistency within groups for patterns of consistency. As can be seen from **Table 4**, the degree of assessment consistency of the unbalanced assessment group tended to improve each time after they assessed a student. Thus, their consistency for Student 1 was 0.80, which then reduced to 0.63 and 0.38 for Students 2 and 3 respectively. The balanced assessment group, despite having the same degree of overall consistency *across* students, demonstrated slight variations *among* students. Their degree of consistency was initially 0.62 for Student 1, then rose to 0.63 for Student 2 before dropping to 0.50 for Student 3. The pattern of consistency of the conflicted assessment teacher was the most unstable and unpredictable with fluctuations at 0.64, 0.42 and 0.54 for Students 1, 2 and 3, respectively.

Reviewing groups' assessment consistency, the teachers in the unbalanced assessment group were one of the two most consistent assessors and their consistency significantly improved across student assessments. The balanced assessment group teachers were more stable in their consistent score assignments, while the conflicted assessment teacher produced the least consistent and most unstable assessments across students.

These results suggest that assessment judgements made by teachers in the conflicted and unbalanced assessment groups are not as reliable as those made by the teachers in the balanced assessment group.

DISCUSSION

Understanding Language Teacher Assessment Decision-Making

This study has aimed "to grasp the process in flight" (Vygotsky and Cole, 1978, p. 68) of teachers' assessment decision-making of students' oral English language skills in the Australian context. The major finding of the study is the identification and confirmation of a three-stage pathway model of teacher language assessment decision-making in which varying strengths of holistic and analytical assessment processes interact to produce one of three final assessment judgement outcomes—balanced, unbalanced or conflicted.

Central to this assessment process is the pivotal role of teacher's first impressions, their judgement Gestalts, that are triggered by initial observations and comparisons of students' language performances. Such Gestalts give a name to the impressionistic, holistic judgements that have received attention in language assessment research (Vaughan, 1991; Mitchell, 1996; Tyndall and Kenyon, 1996; Carr, 2000) as well as in clinical decision-making and other decision-making contexts (Kienle and Kiene, 2011; Cervillin et al., 2014; Danek and Salvi, 2020; Laukkonen et al., 2021) and equate to reported "configurational models of judgement" which are made directly and then checked against specific criteria (Crisp, 2017, p. 35). The findings also confirm the importance of comparative appraisals (Laming, 2004; Heldsinger and Humphry, 2010; Pollitt, 2012; Bartholomew and Yoshikawa, 2018) which naturally arise from serial viewing of student performances and trigger initial judgement Gestalts.

We have seen that how teachers respond to their assessment Gestalt determines the nature and trustworthiness of their final assessment judgement. When teachers engage in robust analysis of task-based assessment criteria to interrogate strong initial "gut feelings," a meta-criterial reframing occurs between holistic and analytical appraisals which enables teachers to arrive at an overall, "on balance" judgement synthesis. When teachers fail to robustly interrogate strong initial Gestalts, it continues as the dominant frame overwhelming analytical processes and results in unbalanced assessment judgements. When teachers engage in fragmented analysis of isolated task criteria in the absence of strong guiding Gestalt, then indecision and conflicted assessment ensues.

These two-way interactions between holistic and analytical judgements highlight the critical role teacher reflexivity and meta-reflection play in sound assessment decision-making. "On balance" judgements may be seen as a "best fit" appraisal with given assessment information (Klenowski et al., 2009, p. 12; Poskitt and Mitchell, 2012, p. 66) characteristic of abductive reasoning (Fischer, 2001). This decision-making synthesis draws on teachers' latent assessment experiences as well as criterion-related assessment information arising from assessment tool engagement, and reflects their meta-criterial interpretations of "the spirit" of assessment rubrics rather than "feature by feature" compliance according to "the letter" (Marshall and Drummond, 2006).

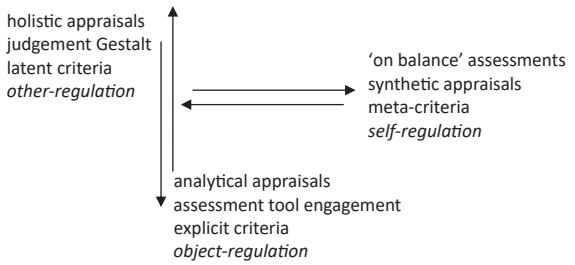


FIGURE 2 | Assessment as dialectic co-regulation.

Allied to this process is the perceived "sufficiency of information" (Smith, 2003) which assessors feel they need in order to make sound decisions. Where there is insufficient information about a student's performance (as is likely in this out-of-context assessment situation), teachers naturally infer, and may even speculate on, contextual information such as student personality traits and behaviours in order to "tip the balance" towards an overall assessment judgement.

Drawing on Frawley's (1987) meditational model of co-regulation and Brookhart's (2016) and Andrade and Brookhart's (2020) co-regulation model of classroom assessment, teacher assessment decision-making can be further theorised as a dialectic process of other- and object-regulation leading to self-regulation, where teachers' final assessment judgements constitute the achievement of a reflexive self-regulated synthesis of holistic and analytical thinking processes. As shown in the meditational model in **Figure 2**, teacher assessment processes involve the dialectic interplay of cognitive regulation arising from perceptions of human others and assessment tool engagement to develop the metacognitive self-regulation of balanced assessment judgements. The model relates these other- object- and self-regulation processes to holistic, analytic and synthetic appraisals aligning them the concepts of latent, explicit and meta-criteria.

The model provides a clearer understanding of the dynamics of each of the assessment pathways. Balanced assessment judgements are the productive self-regulated synthesis of the holism of teachers' assessment Gestalts and the analytics of assessment tool engagement. Unbalanced assessment judgements are the biased outcome of teachers' dominant and insufficiently interrogated assessment Gestalts. Conflicted assessment judgements are the unstable outcome of the unresolved decision-making dilemmas between atomised assessment information from their assessment tool engagement in the absence of a strong guiding assessment Gestalt.

Trustworthiness of Language Teacher Assessment Decision-Making

Teacher assessment decision-making concerns the forming of judgments about the quality of specific performance samples, mediated by assessment resources and the opportunity for teachers to make explicit and justified opinions (Klenowski et al., 2007). Trustworthy assessment has been described as a process where teachers show their disagreements, justify their opinions and arrive at a common, but not necessarily complete, consensus judgement about student performance (Davison, 2004; Davison and Leung, 2009). These notions of assessment trustworthiness are socially anchored in group moderation processes.

Central to the concept of trustworthiness in language assessment are the notions of judgement contestability, process transparency and accountability to evidence. However, these are all key qualities present, or absent in the *individual* dialectic decision-making processes of the three assessment pathways. These pathways show that essential elements of trustworthiness are inherent to the *internal* dynamics of assessment judgement formation. Balanced assessment is trustworthy assessment because it has its own internal self-regulating, self-corrective. In this context, trustworthy assessment can be understood as an internal dialectic process of reflexive co-regulation, in which teachers' final assessment judgements represent a self-regulated decision synthesis of prior holistic and analytical appraisal processes.

The study offers a way forward in understanding and improving the trustworthiness of classroom-based language assessment through a model of how teachers form assessment decision-making judgements. The trustworthiness of unbalanced assessment decision-making is compromised because final assessment judgements are determined by teachers' first perceptions of student performance. Because "perceptions are not reality; perceptions are filtered through the lens that we use to see reality" (Anderson, 2003, p. 145), students' skills are "seen," coloured and constructed through Gestalt's all-encompassing lens. This outcome describes the power of the "halo effect" (Beckwith and Lehmann, 1975; Abikoff et al., 1993; Spear, 1996) where teachers' judgements reflect the extra-performance characteristics of students and unconscious positive or negative biases that threaten assessment trustworthiness.

The halo effect's influence on unbalanced assessment suggests ways it may be remedied to improve its trustworthiness. Teachers' reliance on and confidence in their initial impressions of student performance can minimise the assessment tool engagement and language analysis teachers need to obtain confirming or countervailing information. Alternatively, teachers may engage in tool-mediated language analysis but the strength of their assessment Gestalt based in experience (Barkaoui, 2010a,c, 2011) overrides its influence. In both cases, trustworthiness will be enhanced by the practice of sustained dialogue and meta-reflection within and across the two assessment processes. This remedy is based on the recognition that the strength and quality of teacher reflexivity and interrogation is the key difference between balanced and unbalanced assessment.

The findings on the internal consistency of this pathway group are reassuring. As is evident, the assessment consistency of the unbalanced assessment teachers significantly improved with each assessment of the three students. This shows that the assessment trustworthiness of this group can be readily improved through practice and, as suggested by the literature on resolving unreliable ratings in large-scale testing (Weigle, 1994, 1998;

McNamara, 1996, 2000), should be amenable to training. Given that unbalanced assessment teachers made up the second largest group, such practice effects and training offer the possibility of significant and large-scale improvements in teacher assessment trustworthiness. This example aptly illustrates, at a microgenetic level, how trustworthy expertise develops through repeated practice and quality feedback in stable, regular environments (Kahneman, 2011).

Implications and Possible Future Studies

The study findings add to our understanding of language teacher cognition and assessment literacy underpinning trustworthy language assessment. Identification of assessment decision-making pathways enables diagnosis and correction of judgement errors to enhance the quality of teacher-based assessment. The Gestalt-based assessment decision-making pathway model has practical implications for the content and process of language teacher education. The model can be used in pre-service courses and in-service training as a professional "thinking tool" that enables teachers to view, discuss and understand their thinking processes from an external perspective and to strengthen reflection and metacognition essential for making trustworthy assessment judgements. The study's evidence base for assuring the quality of assessment also strengthens implementation and development of teacher-based assessment policies.

The study also suggests a productive research agenda around the robustness of the model and its applicability to other participants, contexts and language modes and levels. Given that all participants in this study were highly experienced EAL teachers, there is a need to test the model's robustness with less experienced EAL teacher participants such as preservice/beginning/mid-career or untrained EAL teachers. Similarly, as all participants in this study were female, there is a need to examine how well the model reflects the assessment decision-making processes of male teachers. A key issue to be investigated in these studies is what proportion of teachers are found in each assessment pathway group and how these compare with the proportions in this small scale study.

There is also a need to investigate the model with teachers working in different school contexts assessing different language modes of students they already know at different language proficiency levels. For example, the present study could be replicated in relation to trustworthy assessment of student writing (Eckes, 2005, 2008; Barkaoui, 2010b; Leckie and Baird, 2011). In the context of teacher familiarity with students, it would be worth further investigating the influence of any halo effects, for example, in relation to students' personalities or particular language backgrounds. Given the "Out of Context" nature of the study, it would also be worth replicating the study in an "In Context" situation with familiar students known to the teachers. Future studies might also vary the data collection methodology and consider the effectiveness or otherwise of using concurrent, rather than sequential, thinking-aloud protocols in eliciting teachers' assessment thinking.

In view of the documented influence of teacher knowledge, beliefs, expectations and values on their assessment decision-making (McMillan, 2003), there would also be value in investigating how these factors are mobilised before, during and after teacher-based language assessment with a view to improving trustworthiness of teacher assessment. For example, what tacit knowledge of students are reflected in teachers initial assessment Gestalts? What language knowledge is elicited by teachers' use and engagement with assessment tools? What latent criteria do teachers consciously and unconsciously take into account when assessing students' language performances?

Finally, given the insights gained from assessment variability and consistency analysis there would be further value in conducting in-depth, qualitative studies of variability and consistency in teacher assessment decision-making in relation to the three mediational forms of assessment co-regulation. Thus, investigation of tacit, other-regulatory influences of teacher knowledge/perceptions of student characteristics such as gender (Porter and Hang, 1991; O'Loughlin, 2002; Eckes, 2005; Lumley and O'Sullivan, 2005; Ouazad, 2008) and accent (Edwards, 1982; Gass and Varonis, 1984; Gill, 1994; Cargile and Giles, 1998; Major et al., 2002; Carey et al., 2011) and explicit tool-regulatory influences of language assessment tasks (Fayer and Krasinski, 1987; Lumley and McNamara, 1995; McNamara, 1996; Weigle, 1998, 2002; Fulcher and Reiter, 2003; Luoma, 2004; Kim, 2009) and assessment criteria (Weigle, 1999; Lumley, 2002; Rezaei and Lovorn, 2010) would increase our understanding of how these processes interact and combine to produce trustworthy overall assessment judgements according to certain meta-criteria, and suggest new ways to understand and control the sources of teacher assessment variability to improve classroom-based language assessment.

CONCLUSION

The study identified cognitive processes underpinning underexplored teacher-based language assessment decision-making. It empirically established the key role that teachers' first impressions, or assessment Gestalts, play in the formation of assessment judgments and the subsequent interplay between holistic and analytical judgements in three different decision-making pathways. In revealing these pathways, and the Gestalts and factors shaping them, critical issues affecting teacher assessment trustworthiness have been made transparent and can be targeted for remediation and improvement.

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.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the HREA Panel B: Arts, Humanities and Law

of UNSW Australia. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

DP: conceptualisation, professional development program design and implementation, methodology, data collection (facilitated by CD), transcription, data analysis and interpretation, writing—reviewing, and editing. MM: conceptualisation, methodology, data analysis and interpretation, writing—reviewing, and editing. Both authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

Funding from the TEAL Project and support from the Department of Foreign Affairs and Trade is gratefully acknowledged.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2022.830311/full#supplementary-material

REFERENCES

Abikoff, H., Courtney, M., Pelham, W. E., and Koplewicz, H. S. (1993). Teachers' ratings of disruptive behaviors: the influence of halo effects. *J. Abnorm. Child Psychol.* 21, 519–533. doi: 10.1007/BF00916317

Adie, L. E., Klenowski, V., and Wyatt-Smith, C. (2012). Towards an understanding of teacher judgement in the context of social moderation. *Educ. Rev.* 64, 223–240. doi: 10.1080/00131911.2011.598919

Allal, L. (2013). Teachers' professional judgement in assessment: a cognitive act and a socially situated practice. *Assess. Educ.* 20, 20–34.

Alonzo, A. C. (2019). "Defining trustworthiness for teachers' multiple uses of classroom assessment results 1," in *Classroom Assessment and Educational Measurement*, eds S. M. Brookhart, and J. H. McMillan (Oxfordshire: Routledge), 120–145. doi: 10.4324/9780429507533-8

Anderson, L. (2003). *Classroom Assessment: Enhancing The Quality Of Teacher Decision Making*. Oxfordshire: Routledge.

Andrade, H. L., and Brookhart, S. M. (2020). Classroom assessment as the co-regulation of learning. *Assess. Educ.* 27, 350–372.

Baniya, S., Mentzer, N., Bartholomew, S. R., Chesley, A., Moon, C., and Sherman, D. (2019). Using adaptive comparative judgment in writing assessment: an investigation of reliability among interdisciplinary evaluators. *J. Technol. Stud.* 45, 24–35.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: a mixed-method study. *Assess. Writing* 12, 86–107.

Barkaoui, K. (2010a). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Q.* 44, 31–57. doi: 10.5054/tq.2010.214047

Barkaoui, K. (2010c). Variability in ESL essay rating processes: the role of the rating scale and rater experience. *Lang. Assess. Q.* 7, 54–74.

Barkaoui, K. (2010b). Explaining ESL essay holistic scores: a multilevel modeling approach. *Lang. Test.* 27, 515–535.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay rating processes and rater performance. *Assess. Educ.* 18, 279–293. doi: 10.1080/0969594x.2010.526585

Bartholomew, S. R., and Yoshikawa, E. (2018). *A Systematic Review Of Research Around Adaptive Comparative Judgment (ACJ) in K-16 Education. 2018 CTETE Monograph Series*. Blacksburg, VA: Virginia Polytechnic Institute and State University.

Beckwith, N. E., and Lehmann, D. R. (1975). The importance of halo effects in multi-attribute attitude models. *J. Market. Res.* 12, 265–275.

Black, P., and Wiliam, D. (1998). Assessment and classroom learning. *Assess. Educ.* 5, 7–74. doi: 10.4324/9781315123127-3

Borg, S. (2009). *Language Teacher Cognition. The Cambridge Guide To Second Language Teacher Education*. Cambridge, MA: Cambridge University Press, 163–171.

Bowers, V. A., and Snyder, H. L. (1990). Concurrent versus retrospective verbal protocol for comparing window usability. *Proc. Hum. Fact. Soc. Annu. Meet.* 34, 1270–1274.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educ. Meas.* 22, 5–12.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educ. Meas.* 30, 3–12. doi: 10.1111/j.1745-3992.2010.00195.x

Brookhart, S. M. (2016). "Section discussion: Building assessments that work in classrooms," in *Handbook of Human And Social Conditions In Assessment*, eds G. T. L. Brown and L. R. Harris (New York, NY: Routledge), 351–365.

Carey, M. D., Mannell, R. H., and Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Lang. Test.* 28, 201–219.

Cargile, A. C., and Giles, H. (1998). Language attitudes toward varieties of English: an American-Japanese context. *J. Appl. Commun. Res.* 26, 338–356.

Carless, D. (2007). Learning-oriented assessment: conceptual bases and practical implications. *Innov. Educ. Teach. Int.* 44, 57–66.

Carr, N. T. (2000). A comparison of the effects of analytic and holistic rating scale types in the context of composition tests. *Issues Appl. Linguist.* 11, 207–241.

Castleton, G., Wyatt-Smith, C., Cooksey, R., and Freebody, P. (2003). The nature of teachers' qualitative judgements: a matter of context and salience: part two: out-of-context judgements. *Aust. J. Lang. Lit.* 26, 33–42.

Cervillin, G., Borghi, L., and Lippi, G. (2014). Do clinicians decide relying primarily on Bayesians principles or on Gestalt perception? Some pearls and pitfalls of Gestalt perception in medicine. *Int. Emergen. Med.* 9, 513–519. doi: 10.1007/s11739-014-1049-8

Clarà, M. (2014). Understanding teacher knowledge from a cultural psychology approach. *Teach. Teach. Educ.* 43, 110–119.

Clark, C. M., and Peterson, P. L. (1984). "Teachers' Thought Processes. Occasional Paper No. 72," in *Handbook of Research on Teaching*, Third Edn, ed. M. C. Wittrock (New York, NY: Macmillan).

Cooksey, R. W. (1996). *Judgment Analysis: Theory, Methods, And Applications*. Cambridge, MA: Academic press.

Cooksey, R. W., Freebody, P., and Wyatt-Smith, C. (2007). Assessment as judgment-in-context: analysing how teachers evaluate students' writing 1. *Educ. Res. Eval.* 13, 401–434.

Cresswell, M. J. (1997). *Examining Judgements: Theory And Practice Of Awarding Public Examination Grades*, Doctoral dissertation, Institute of Education, University of London.

Crisp, V. (2010). Judging the grade: exploring the judgement processes involved in examination grading decisions. *Evaluation Res. Educ.* 23, 19–35.

Crisp, V. (2013). Criteria, comparison and past experiences: how do teachers make judgements when marking coursework? *Assess. Educ.* 20, 127–144.

Crisp, V. (2017). The judgement processes involved in the moderation of teacher-assessed projects. *Oxford Rev. Educ.* 43, 19–37.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Lang. Test.* 7, 31–51. doi: 10.1177/026553229000700104

Cumming, A. (2002). Assessing L2 writing: alternative constructs and ethical dilemmas. *Assess. Writ.* 8, 73–83. doi: 10.1016/S1075-2935(02)00047-8

Cumming, A., Kantor, R., and Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: a descriptive framework. *Modern Lang. J.* 86, 67–96. doi: 10.1111/1540-4781.00137

Danek, A. H., and Salvi, C. (2020). Moment of truth: why Aha! experiences are correct. *J. Creat. Behav.* 54, 484–486.

Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools. *Lang. Test.* 21, 305–334.

Davison, C. (2008). "Assessment for learning: Building inquiry-oriented assessment communities," in *Proceedings of the 42nd Annual TESOL Convention and Exhibit*, New York.

Davison, C. (2013). "Innovation in assessment: Common misconceptions and problems," in *Innovation and change in English language education*, eds K. Hyland, and L. L. C. Wong (Milton Park: Routledge), 279–292.

Davison, C. (2017). "Enhancing teacher assessment literacy in English language education: Problems and pitfalls," in *Proceedings of the Plenary presented at the applied linguistics conference (ALANZ/ALAA/ALTAANZ)*, Auckland.

Davison, C. (2019). "*Using Assessment To Enhance Learning In English Language Education*," in *Second Handbook of English Language Teaching*, ed. X. Gao (New York, NY: Springer), 433–454. doi: 10.1007/978-3-030-02899-2_21

Davison, C. (ed.) (2021). "Enhancing teacher assessment literacy: one approach to improving teacher knowledge and skills in Australia," in *Envisioning Teaching and Learning of Teachers for Excellence and Equity in Education*, (Singapore: Springer), 33–43.

Davison, C., and Leung, C. (2009). Current issues in english language teacher-based assessment. *TESOL Q.* 43, 393–415.

Davison, C., and Michell, M. (2014). EAL assessment: what do Australian teachers want? *TESOL Context* 24, 51–72.

Dochy, F. (2009). *Assessment, Learning And Judgement In Higher Education*. New York, NY: Springer, 1–30.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: a many-facet Rasch analysis. *Lang. Assess .Q.* 2, 197–221.

Eckes, T. (2008). Rater types in writing performance assessments: a classification approach to rater variability. *Lang. Test.* 25, 155–185. doi: 10.1177/0265532207086780

Edwards, J. R. (1982). *Language Attitudes And Their Implications Among English Speakers. Attitudes Toward Language Variation*. Milton Park: Routledge, 20–33.

Eisner, E. W. (1998). *Educational Connoisseurship. The Enlightened Eye, USA*. Hoboken, NJ: Prentice Hall.

Esterberg, K. G. (2002). *Qualitative Methods in Social Research*. New York, NY: McGraw–Hill.

Fayer, J. M., and Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Lang. Learn.* 37, 313–326.

Feryok, A., and Pryde, M. (2012). Images as orienting activity: using theory to inform classroom practices. *Teach. Teach.* 18, 441–454.

Fischer, H. R. (2001). Abductive reasoning as a way of worldmaking. *Found. Sci.* 6, 361–383.

Frawley, W. J. (1987). *Text And Epistemology*. Norwood. NJ: Ablex.

Freeman, D. (2002). The hidden side of the work: teacher knowledge and learning to teach. A perspective from North American educational research on teacher education in English language teaching. *Lang. Teach.* 35, 1–13.

Fulcher, G., and Reiter, R. M. (2003). Task difficulty in speaking tests. *Lang. Test.* 20, 321–344. doi: 10.1191/0265532203lt259oa

Gamaroff, R. (2000). Rater reliability in language assessment: the bug of all bears. *System* 28, 31–53.

Gass, S., and Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Lang. Learn.* 34, 65–87.

Gill, M. M. (1994). Accent and stereotypes: their effect on perceptions of teachers and lecture comprehension. *J. Appl. Commun. Res.* 22, 348–361. doi: 10.1080/00909889409365409

Gipps, C. (2012). *Beyond Testing: Towards a Theory Of Educational Assessment*. London, UK: Falmer Press.

Glaser, B. G., and Strauss, A. (1967). *The Discovery Of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine.

Glogger-Frey, I., Herppich, S., and Seidel, T. (2018). Linking teachers' professional knowledge and teachers' actions: judgment processes, judgments and training. *Teach. Teach. Educ.* 76, 176–180.

Good, T. L., and Lavigne, A. L. (2017). *Looking in Classrooms*. New York, NY: Routledge.

Hamp-Lyons, L. (ed.) (1991). *Assessing Second Language Writing in Academic Contexts*. Norwood, NJ: Ablex Publishing Corporation.

Harlen, W. (2005). Trusting teachers' judgement: research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Res. Pap. Educ.* 20, 245–270. doi: 10.1080/02671520500193744

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

Huot, B. (1990). The literature of direct writing assessment: major concerns and prevailing trends. *Rev. Educ. Res.* 60, 237–263. doi: 10.3102/00346543060002237

Jäkel, F., Singh, M., Wichmann, F. A., and Herzog, M. H. (2016). An overview of quantitative approaches in Gestalt perception. *Vision Res.* 126, 3–8. doi: 10.1016/j.visres.2016.06.004

Janopoulos, M. (1993). "Comprehension, communicative competence, and construct validity: holistic scoring from an ESL perspective," in *Validating Holistic Scoring for Writing Assessment*, eds M. W. Williamson and B. A. Huot (Cresskill, NJ: Hampton Press), 303–322.

Johnson, B., and Christensen, L. (2010). *Educational Research: Quantitative, Qualitative, And Mixed Approaches*. Thousand Oaks, CA: Sage Publications.

Joughin, G. (2009a). *Introduction: Refocusing Assessment. In Assessment, Learning And Judgement In Higher Education*. Dordrecht: Springer, 1–11.

Joughin, G. (2009b). *Assessment, Learning And Judgement In Higher Education: A Critical Review. Assessment, Learning And Judgement In Higher Education*. Dordrecht: Springer, 13–27.

Kahneman, D. (2011). *Thinking, Fast And Slow*. Basingstoke: Macmillan.

Kienle, G. S., and Kiene, H. (2011). Clinical judgement and the medical profession. *J. Eval. Clin. Pract.* 17, 621–627. doi: 10.1111/j.1365-2753.2010.01560.x

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral english performance: a mixed methods approach. *Lang. Test.* 26, 187–217.

Klein, G. (1997). "The recognition-primed decision (RPD) model: looking back, looking forward," in *Naturalistic Decision Making*, eds C. E. Zsambok and G. Klein (Mahwah, NJ: Lawrence Erlbaum Associates), 285–292.

Klenowski, V. (2013). Towards improving public understanding of judgement practice in standards-referenced assessment: an Australian perspective. *Oxford Rev. Educ.* 39, 36–51.

Klenowski, V., Adie, L., Gunn, S., Looney, A., Elwood, J., Wyatt-Smith, C., et al. (2007). "Moderation as judgement practice: reconciling system level accountability and local level practice," in *Proceedings of the Australian Association for Research in Education 2007 Conference: Research impacts: Proving or improving?* (Melbourne VIC: Australian Association for Research in Education), 1–29.

Klenowski, V., Adie, L., Gunn, S., Looney, A., Elwood, J., Wyatt-Smith, C., et al. (2009). Moderation as judgement practice: reconciling system level accountability and local level practice. *Curr. Perspect.* 29, 10–28.

Klenowski, V., and Wyatt-Smith, C. (2010). Standards, teacher judgement and moderation in contexts of national curriculum and assessment reform. *Assess. Matters* 2, 107–131. doi: 10.18296/am.0078

Knight, P. (2007). "Grading, classifying and future learning," in *Rethinking Assessment In Higher Education*, eds D. Boud, and N. Falchikov (Milton Park: Routledge), 82–96. doi: 10.4324/9780203964309-14

Koffka, K. (1922). Perception: and introduction to the Gestalt-theorie. *Psychol. Bull.* 19, 531–585. doi: 10.1037/h0072422

Koffka, K. (1935). *Principles of Gestalt Psychology*. London: Lund Humpries.

Koffka, K. (2013). *Principles of Gestalt Psychology*, Vol. 44. New York, NY: Routledge.

Koh, K. H. (2011). Improving teachers' assessment literacy through professional development. *Teach. Educ.* 22, 255–276. doi: 10.1080/10476210.2011.593164

Korthagen, F. A. (2010). Situated learning theory and the pedagogy of teacher education: towards an integrative view of teacher behavior and teacher learning. *Teach. Teach. Educ.* 26, 98–106. doi: 10.1016/j.tate.2009.05.001

Kubanyiova, M., and Feryok, A. (2015). Language teacher cognition in applied linguistics research: revisiting the territory, redrawing the boundaries, reclaiming the relevance. *Modern Lang. J.* 99, 435–449.

Lado, R. (1961). *Language Testing: The Construction and Use of Foreign Language Tests. A Teacher's Book*. New York, NY: McGraw-Hill.

Laming, D. (2004). *Human Judgement The Eye Of The Beholder*. London: Thompson Learning.

Laukkonen, R. E., Ingledew, D. J., Grimmer, H. J., Schooler, J. W., and Tangen, J. M. (2021). Getting a grip on insight: real-time and embodied Aha experiences predict correct solutions. *Cogn. Emot.* 35, 918–935. doi: 10.1080/02699931.2021.1908230

Laukkonen, R. E., Schooler, J. W., and Tangen, J. M. (2018). *The Eureka Heuristic: Relying On Insight To Appraise The Quality Of Ideas*. Berlin: Springer.

Leckie, G., and Baird, J. A. (2011). Rater effects on essay scoring: a multilevel analysis of severity drift, central tendency, and rater experience. *J. Educ. Meas.* 48, 399–418. doi: 10.1111/j.1745-3984.2011.00152.x

Leung, C. (2013). "Classroom-based assessment issues for language teacher education," in *The Companion To Language Assessment*, Vol. 3, ed. A. J. Kunnan (Hoboken, NJ: Wiley & Sons), 1510–1519. doi: 10.1002/9781118411360.wbcla064

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Lang. Test.* 19, 246–276. doi: 10.1191/0265532202lt230oa

Lumley, T., and McNamara, T. (1995). Rater characteristics and rater bias: implications for training. *Lang. Test.* 12, 54–71. doi: 10.1177/026553229501200104

Lumley, T., and O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Lang. Test.* 22, 415–437.

Luoma, S. (2004). *Assessing Speaking: Ernst Klett Sprachen*. Cambridge: Cambridge University Press.

Major, R. C., Fitzmaurice, S. F., Bunta, F., and Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: implications for ESL assessment. *TESOL Q.* 36, 173–190. doi: 10.2307/3588329

Marshall, B., and Drummond, M. (2006). How teachers engage with assessment for learning: lessons from the classroom. *Res. Pap. Educ.* 21, 133–149.

McMahon, S., and Jones, I. (2015). A comparative judgement approach to teacher assessment. *Assess. Educ.* 22, 368–389.

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: implications for theory and practice. *Educ. Meas.* 22, 34–43. doi: 10.1111/j.1745-3992.2003.tb00142.x

McMillan, J. H., and Nash, S. (2000). "Teacher Classroom Assessment and Grading Practices Decision Making," in *Paper Presented at the Annual Meeting of the National Council on Measurement in Education*, New Orleans, LA.

McNamara, T. (1996). *Measuring Second Language Performance*. London: Longman.

McNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press.

Mertler, C. A. (2004). Secondary teachers' assessment literacy: does classroom experience make a difference? *Am. Second. Educ.* 43, 49–64.

Michell, M., and Davison, C. (2020). "Bringing the teacher back in': toward l2 assessment praxis in english as an additional language education," in *Toward a Reconceptualization of Second Language Classroom Assessment*, eds M. E. Poehner, and O. Inbar-Lourie (Cham: Springer), 23–41.

Miles, M. B., and Huberman, A. M. (1994). *Qualitative Data Analysis: An Expanded Sourcebook*. Thousand Oaks, CA: SAGE.

Mitchell, S. E. (1996). Institutions, individuals and talk: the construction of identity in fine art. *Int. J. Art Design Educ.* 15, 143–154. doi: 10.1111/j.1476-8070.1996.tb00661.x

Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educ. Meas.* 22, 13–25. doi: 10.1111/j.1745-3992.2003.tb00140.x

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assess. Educ.* 14, 149–170. doi: 10.1080/09695940701478321

Nunes, A. K. F., Barroso, R. D. C. A., and Santos, J. F. (2019). "The use of Triangulation as a tool for validation of data in qualitative research in Education," in *Proceedings of the World Conference on Qualitative Research*, Portugal, Vol. 1, 334–336.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Lang. Test.* 19, 169–192. doi: 10.1191/0265532202lt226oa

Ouazad, A. (2008). Assessed by a Teacher Like Me: Race, Gender, And Subjective Evaluations. International Journal of Art & Design Education (INSEAD Working Paper No. 2008/57/EPS). Available online at: Retrieved from https://ssrn.com/abstract=1267109 (accessed October 27, 2016).

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assess. Educ.* 19, 281–300.

Popham, J. W. (2014). *Classroom Assessment: What Teachers Need To Know*. London: Pearson.

Popham, W. J. (2004). Why assessment illiteracy is professional suicide. *Educ. Leadersh.* 62:82.

Popham, W. J. (2009). Assessment literacy for teachers: faddish orfundamental? *Theory Pract.* 48, 4–11.

Porter, D., and Hang, S. S. (1991). Sex, status and style in the interview. *Dolphin* 21, 117–128.

Poskitt, J., and Mitchell, K. (2012). New Zealand teachers' overall teacher judgements (OTJs): equivocal or unequivocal? *Assess. Matters* 4, 53–75.

Prasad, G. R. (2021). Enhancing clinical judgement in virtual care for complex chronic disease. *J. Eval. Clin. Pract.* 27, 677–683. doi: 10.1111/jep.13544

Rea-Dickins, P. (2004). Understanding teachers as agents of assessment. *Lang. Test.* 21, 249–258.

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

Rosch, E. (1978). *Principles of Categorization Text. Cognition and Categorization*. Hillsdale, NJ: Lawrence Erlbaum Associates, 24.

Sadler, D. R. (1985). The origins and functions of evaluative criteria. *Educ. Theory* 35, 285–297. doi: 10.1111/j.1741-5446.1985.00285.x

Sadler, D. R. (1998). Formative assessment: revisiting the territory. *Assess. Educ.* 5, 77–84.

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 (Dordrecht: Springer), 1–19. doi: 10.1007/978-1-4020-8905-3_4

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educ. Res.* 29, 4–14.

Shepard, L. A. (2001). "The role of classroom assessment in teaching and learning," in *Handbook of Research On Teaching*, ed. V. Richardson (Washington, D.C: AERA).

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educ. Meas.* 22, 26–33. doi: 10.1111/j.1745-3992.2003.tb00141.x

Spear, M. (1996). The influence of halo effects upon teachers' assessments of written work. *Res. Educ.* 56, 85–87.

Stiggins, R. J. (2002). Assessment crisis: the absence of assessment for learning. *Phi Delta Kappan* 83, 758–765. doi: 10.1177/003172170208301010

Stiggins, R. J., Arter, J. A., Chappuis, J., and Chappuis, S. (2004). *Classroom Assessment For Student Learning: Doing it right–using it well*. Portland: Assessment Training Institute.

Taras, M. (2009). Summative assessment: the missing link for formative assessment. *J. Furth. High. Educ.* 33, 57–69. doi: 10.1080/03098770802638671

Taylor, L. (2006). The changing landscape of english: implications for language assessment. *ELT J.* 60, 51–60. doi: 10.1093/elt/cci081

Taylor, L. (2009). Developing assessment literacy. *Annu. Rev. Appl. Linguist.* 29, 21–36.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Lang. Learn.* 44, 307–336. doi: 10.1111/j.1467-1770.1994.tb01104.x

Turner, C. E., and Purpura, J. E. (eds) (2016). "16. Learning-oriented assessment in second and foreign language classrooms," in *Handbook Of Second Language Assessment*, (Berlin: De Gruyter Mouton), 255–274.

Tyndall, B., and Kenyon, D. M. (1996). "Validation of a new holistic rating scale using Rasch multifaceted analysis," in *Validation in Language Testing*, eds A. H. Cumming and R. Berwick (Clevedon, UK: Multilingual Matters), 39–57.

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

Van Den Haak, M., De Jong, M., and Schellens, P. J. (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behav. Inform. Technol.* 22, 339–351. doi: 10.1080/0044929031000

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 Publishing Corporation), 111–125.

Vygotsky, L. S., and Cole, M. (1978). *Mind in society: Development Of Higher Psychological Processes*. Cambridge, MA: Harvard university press.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., et al. (2012a). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychol. Bull.* 138, 1172–1217. doi: 10.1037/a0029333

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., van der Helm, P. A., et al. (2012b). A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychol. Bull.* 138:218. doi: 10.1037/a0029334

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Lang. Test.* 11, 197–223. doi: 10.1177/026553229401100206

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Lang. Test.* 15, 263–287.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: quantitative and qualitative approaches. *Assess. Writ.* 6, 145–178. doi: 10.1016/S1075-2935(00)00010-6

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Wertheimer, M. (1912). Experimental studies on the seeing of motion. *Psychologia* 61, 161–165.

Wertheimer, M. (1923). "Laws of organization in perceptual forms," in *A Source Book of Gestalt Psychology*, ed. W. D. Ellis (London: Routledge), 7188.

Wertheimer, M. (1938). "The general theoretical situation," in *A source book of Gestalt psychology*, ed. W. D. Ellis (London: Routledge & Kegan Paul), 12–16.

Wertheimer, M. (2012). "Experimental studies on seeing motion," in *On Perceived Motion And Figural Organization*, ed. L. Spillmann (Cambridge, MA: The MIT Press), 1–92.

Wilen, W., Bosse, M. I., Hutchison, J., and Kindsvatter, R. (2004). *Dynamics of Effective Secondary Teaching*, 5th Edn. Boston, MA: Allyn and Bacon.

Williamson, M. M., and Huot, B. A. (1993). *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations.* Cresskill, NJ: Hampton Press.

Wyatt-Smith, C., and Adie, L. (2021). The development of students' evaluative expertise: enabling conditions for integrating criteria into pedagogic practice. *J. Curr. Stud.* 53, 399–419.

Wyatt-Smith, C., and Klenowski, V. (2013). Explicit, latent and meta-criteria: types of criteria at play in professional judgement practice. *Assess. Educ.* 20, 35–52. doi: 10.1080/0969594x.2012.725030

Wyatt-Smith, C., and Klenowski, V. (2014). "Elements of better assessment for the improvement of learning," in *Designing Assessment For Quality Learning*, ed. E. Wyatt-Smith (Dordrecht: Springer), 195–210. doi: 10.1007/978-94-007-5902-2_13

Wyatt-Smith, C., Castleton, G., Freebody, P., and Cooksey, R. (2003). The nature of teachers' qualitative judgements: a matter of context and salience: part one:'In-context'judgement. *Aust. J. Lang. Literacy* 26, 11–32.

Wyatt-Smith, C., Klenowski, V., and Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: a study of standards in moderation. *Assess. Educ.* 17, 59–75.

Xu, Y., and Brown, G. T. (2016). Teacher assessment literacy in practice: a reconceptualization. *Teach. Teach. Educ.* 58, 149–162.

Yin, M. (2010). Understanding classroom language assessment through teacher thinking research. *Lang. Assess. Q.* 7, 175–194. doi: 10.1080/15434300903447736

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.

The handling editor declared a shared affiliation, though no other collaboration, with one of the author MM at the time of the review.

frontiers | Frontiers in Education

ORIGINAL RESEARCH
published: 28 June 2022
doi: 10.3389/feduc.2022.925352

Supporting Teachers in Improving Formative Decision-Making: Design Principles for Formative Assessment Plans

Janneke van der Steen[1,2], Tamara van Schilt-Mol[2], Cees van der Vleuten[1] and Desirée Joosten-ten Brinke[3]*

[1] Faculty of Health, Medicine and Life Sciences, School of Health Professions Education, Maastricht University, Maastricht, Netherlands, [2] Research Center for Contemporary Assessment and Decision Making, School for Education, HAN University of Applied Sciences, Nijmegen, Netherlands, [3] Department of Online Learning and Instruction, Faculty of Educational Sciences, Open Universiteit, Heerlen, Netherlands

OPEN ACCESS

Edited by:
Chris Davison, University of New South Wales, Australia

Reviewed by:
Serafina Pastore, University of Bari Aldo Moro, Italy
Zhengdong Gan, University of Macau, China

***Correspondence:**
Janneke van der Steen
Janneke.vandersteen@maastrichtuniversity.nl

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

Received: *21 April 2022*
Accepted: *07 June 2022*
Published: *28 June 2022*

Citation:
van der Steen J, van Schilt-Mol T, van der Vleuten C and Joosten-ten Brinke D (2022) Supporting Teachers in Improving Formative Decision-Making: Design Principles for Formative Assessment Plans. Front. Educ. 7:925352. doi: 10.3389/feduc.2022.925352

Formative assessment is considered as one of the most effective interventions to support teacher decision-making and improve education and student learning. However, formative assessment does not always meet these expectations. In order to be effective, formative assessment activities should be consciously and coherently planned aligned with other aspects of the curriculum and the decisions teachers wish to make based on these activities. While there is sufficient support for teachers to design formative assessment activities, no guidelines exist to help them tie these different activities together in an effective way. To support teachers in designing formative assessment plans informing formative decision-making, this study focused on the creation of a set of design principles. These design principles for formative assessment plans were formulated based on expert interviews and subsequently evaluated by future users. The result is a set of eight design principles that can be used and validated in educational practice.

Keywords: evaluation utilization, formative assessment, design principles, teacher decision-making, classroom assessment

INTRODUCTION

Assessment is used formatively when teachers and/or students interpret and use the evidence about student achievement to make formative decisions, decisions about the next steps in teaching and learning (Black and Wiliam, 2009). For example, decisions about adjusting lessons, how to differentiate, if students are ready for a new subject or what is the best way to support student learning at a given time. Accordingly, formative assessment embodies all activities that students and teachers undertake to elicit evidence to establish where students are in their learning in order to inform education. Teachers interpret and use this information to "make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of this information" (Black and Wiliam, 2009, p. 9).

In fact, formative assessment is a continuous dialogue between students and teachers about three questions (Wiliam and Thompson, 2007; Black and Wiliam, 2009; Wiliam, 2011):

1) Where is the learner going?
2) Where is the learner now?
3) What is necessary to bridge the gap between where the learners are in their learning and where they are going?

Many studies have presented strategies that help to answer these three questions. Strategies associated with formative assessment include: identifying and making explicit learning objectives and success criteria; elicitation of evidence of students' understanding or learning; interpretation of the elicited information against the learning objectives and/or success criteria; providing students with feedback, and follow-up actions taken by the student and/or teacher to improve teaching and learning (Ruiz-Primo and Furtak, 2007; Antoniou and James, 2014; Veugen et al., 2021). Continuously answering these questions, using these strategies, helps teachers to better meet students' needs and to increase students' involvement in their own learning process (Black and Wiliam, 2010). As a result, formative assessment is seen as one of the most effective interventions to improve education and increase student learning (Briggs et al., 2012; Christoforidou et al., 2014; Offerdahl et al., 2018).

Formative assessment has an intuitive appeal and the potential effectiveness is widely acknowledged (Furtak and Ruiz-Primo, 2008; Black and Wiliam, 2009; Offerdahl et al., 2018), nevertheless empirical evidence about the effect of formative assessment on student learning is variable. Empirical studies that investigated the effect sizes of formative assessment on student learning vary in methods and outcomes (Black and Wiliam, 2010; Kingston and Nash, 2011; Briggs et al., 2012; Offerdahl et al., 2018; Gu, 2021). Offerdahl et al. (2018) suggest that differences in enactment by teachers explain a main part of the differences in effect sizes of formative assessment on student learning. Perspectives on formative assessment, context or formative assessment proficiency and literacy can all influence this enactment (Deneen et al., 2019; Earle, 2021; Gu, 2021; Yan et al., 2021). Apart from these factors improving enactment of formative assessment starts with answering the question how teachers can best design and implement formative assessment to have it really contribute to better or better founded formative decision-making. This article will try to answer this question by focusing on teacher activities in designing and implementing formative assessment.

Formative assessment is best considered as an ongoing process to inform and support teaching and learning (Earle, 2021; Gu, 2021; Veugen et al., 2021). While achieving learning objectives usually exceeds a lesson, teachers working with formative assessment also need to exceed lessons planning these activities. During a series of lessons they have to keep checking whether objectives are reached or not and for what reasons, followed by deciding what this means for their teaching. When teachers want to design formative assessment, they, therefore, should plan series of connected formative activities instead of individual activities to support their lessons (Furtak et al., 2016). These formative activities should be constructively aligned with the objectives and planned lessons. Especially this connection between formative assessment activities and the link with the rest of the curriculum seems important but also hard to accomplish in classroom practice. Many studies that investigated formative assessment conclude that extra attention for the integration, coherency and alignment of formative assessment activities in classroom practice is needed to be more effective (Gulikers et al., 2013; Wylie and Lyon, 2015; Van Den Berg, 2018).

For planning these connected series of formative assessment activities, Wiliam (2013, 2014) advocates decision-driven data collection. In decision-driven data collection future formative decisions are the starting point for planning formative assessment activities (Wiliam, 2013, 2014; Moss, 2020). It differs from one of the most well-known forms of formative assessment, namely data-based decision making (Schildkamp and Kuiper, 2010; Van der Kleij et al., 2015; Heitink et al., 2016). Since data-based decision making starts from existing data, Wiliam (2013) argues that this might, in certain situations, be unsuitable or just too late for formative decision-making about for instance lesson preparations or last-minute adjustments in instruction. Accordingly, he suggests making plans of actions, a blueprint of formative learning activities, that incorporate the strategies for collecting evidence of learning as well as what will be done with this information when it is collected based on future formative decisions (Wiliam, 2013). Formative assessment plans for decision-driven data collection, as suggested by Wiliam (2013), can also accommodate the integration, coherency and alignment in advance to support teachers in implementing formative assessment that informs their formative decision-making. So far, formative assessment, however, has predominantly been planned, executed and investigated in singular formative assessment activities. As a result, we see a lot of examples and tools for teachers to design formative assessment activities but few examples or guidelines to help them tie these different activities together in an effective way in formative assessment plans.

Hence, the research question in this study is:

What are design principles for formative assessment plans that help teachers to make better founded formative decisions in classroom practice about the next steps in teaching and learning?

MATERIALS AND METHODS

This study has an educational design research approach and intends to develop a first prototype set of design principles for future use and empirical testing in schools for secondary education. For this design study, the three steps for educational design research, defined by McKenney and Reeves (2019), were followed (see **Figure 1**).

In the first step, analysis and exploration, interviews with three different groups of experts were used to gather the first ideas about design principles for formative assessment plans. Subsequently, in the design and construction phase, these expert interviews were followed by a thematic analysis of the interview data to design and construct a set of design principles. Finally, as part of the concluding step of evaluation and reflection, future users evaluated the constructed design principles. Teachers from four schools for secondary education evaluated the design principles during group interviews, which resulted in the final adjustments of the design principles.

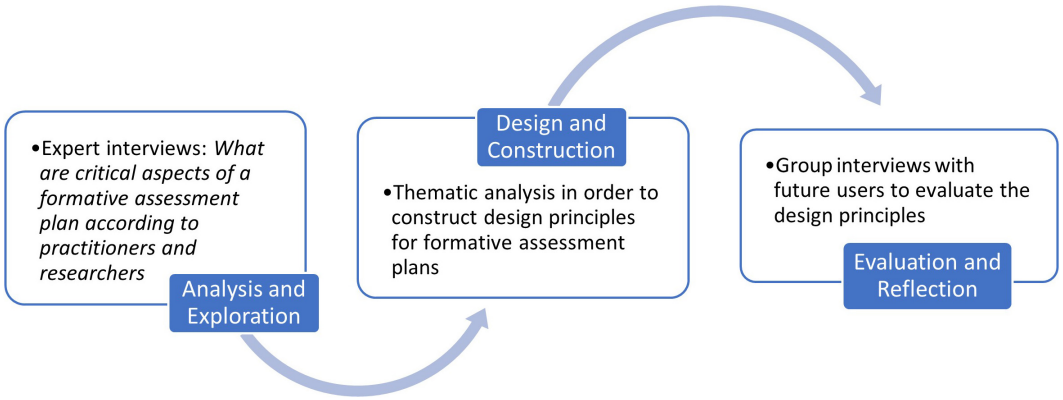


FIGURE 1 | Subsequent research-activities presented through steps of educational design research (McKenney and Reeves, 2019).

To uphold the participative character of educational design research, future users were involved as members of the expert groups and as informants in the evaluation phase. As experiential experts in the expert group interviews, they helped constructing the first ideas about design principles for formative assessment plans and made a first step in analyzing the outcomes. In the evaluation phase, future users were asked to evaluate the design principles.

Analysis and Exploration: Expert Interviews

Three interviews were planned with heterogeneous groups of experts on the subject formative assessment. In total, twenty participants were experts from both research and practice, all involved with formative assessment. The educational researchers were selected as experts when they conducted research on formative assessment but also had their own formative assessment teaching practice. Teachers were selected as experts when they worked in one of four participating schools and had demonstrable experience using formative assessment in their classrooms. One school-policy maker was selected because she was responsible for implementing formative assessment in one of these four schools. Additionally, two teacher educators were selected who were involved in the development of a minor for formative assessment. **Table 1** shows the combination of experts in each expert group.

The purpose of these expert interviews was to reach agreement among the participants of each group about what they thought were critical aspects a formative assessment plan should have to be effective. These critical aspects will be clustered, first within groups and then across groups, and used in a later stage as a starting point to formulate design principles. To promote consensus, the interviews were organized as group decision rooms where the discussion was supported by a digital group support system (Fjermestad and Hiltz, 2000; Pyrko et al., 2019). This support system offers participants the possibility to answer questions individually and digitally through a device followed by a group discussion about how to cluster all given answers. By clustering their own answers during the interview within the group, participants were directly involved in the first phase of data-analysis. All subsequent steps that were taken in the expert interviews are presented in **Table 2**.

TABLE 1 | Participants expert interviews.

Groups	Participants
Group 1	Three educational researchers, one teacher-educator, and one school policy maker
Group 2	Two educational researchers, three teachers from one school for secondary education
Group 3	Nine teachers from two school for secondary education, one teacher-educator

Through these steps, each expert interview has generated a list of clusters of critical aspects for formative assessment plans. Going forward in this article, these clusters of aspects will be referred to as features of formative assessment plans.

Design and Construction: Thematic Analysis

The features of formative assessment plans that were suggested through the expert interviews were collected and used in a design session with three of the four authors of this paper and two of their colleague researchers. The purpose of this design session was to synthesize the outcomes from the three expert group interviews and develop a first draft of design principles for formative assessment plans. Thematic analysis is systematic but always subjective (Bearman and Dawson, 2013). Therefore, the researchers who joined the thematic analysis during this design session were invited since they all had experience with investigating formative assessment, knew the project well but operated at different levels of distance in the project. Intersubjectivity was sought by combining the common knowledge of the objectives in the current study with the quality of these researchers with different backgrounds, experiences, and perspectives on formative assessment. To systematically generate design principles from the collected interview data the first five phases of thematic analysis as described by Nowell et al. (2017) were followed during this session. At the start, to get familiarized

TABLE 2 | Activities and questions expert interviews.

Activity	Question
1. Participants answer individually	"Can you name three critical aspects of a formative assessment plan?"
2. Group discussion	"How can we cluster the given aspects? What name should the different clusters have?"
3. Participants answer individually	"Which two critical aspects a formative assessment plan should have are still missing in the composed list?"
4. Group discussion	"Can we add these extra aspects to existing clusters or do we need to create new ones?"
5. Participants answer individually	"If you still think that there are critical aspects missing in the composed list, can you please add them now?"
6. Group discussion	"Can we add these extra aspects to existing clusters or do we need to create new ones?"
7. Participants answer individually	"How would you arrange all clustered critical aspects that are the result of this expert interview, in order of importance?"

with all data, each feature, that was a result of the expert interviews, was put on an individual card and these cards were laid down on a large table. The researchers studied these cards, clustered the features that were similar into themes, and named each of these themes. Thereafter, in the next phase of analysis, these themes were critically reviewed by the researchers by questioning if each theme of features was explicitly applicable for designing formative assessment plans. This final critical review resulted in 10 final themes of features of formative assessment plans. These remaining 10 themes were then formulated as design principles, in phase five, using the following structure based on Van den Akker's (2013) suggestion on how to formulate design principles:

If you want to design formative assessment plans

- *Then you are best advised to give these plans the following characteristics*

These design principles were provided with a description of what this would mean in practice and used as input for the group interviews with future users. In these group interviews, future users will evaluate this draft version of the design principles for formative assessment plans based on transparency, completeness, usability, and suitability for teaching practice.

Evaluation and Reflection: Group Interviews

Four group interviews were set up to evaluate the draft version of the design principles for formative assessment plans. The group interviews were organized with future users originating from four schools for secondary education. Each group consisted of five to eight teachers from the same school. In two cases, a school leader also joined the interview (see **Table 3**).

The teachers and school leaders were questioned about recommendations regarding transparency, usability, completeness, and suitability of the design principles for school practice. The participants had received the design principles in advance.

TABLE 3 | Participants group interviews.

Groups	Participants
School 1	Five teachers
School 2	Seven teachers and two school leaders
School 3	Five teachers and two school leaders
School 4	Six teachers and one school policy maker

First, the participants were asked to write down all recommendations they could think of to improve the design principles. Secondly, they were asked to decide what facet of the design principles would improve if this recommendation was followed. Facets they could choose from were transparency, usability, completeness, or suitability. Subsequently, they were asked to give explanations of their recommendations and the improvements they would expect. The interview transcripts were analyzed through thematic analysis (Nowell et al., 2017). Recommendations for improvements from the interviews were coded and clustered into themes by the first author. Before defining and naming them, three of the four authors reviewed the initial themes and subthemes. Each theme of recommendations was also linked to the corresponding facets of the design principles that would improve most when the recommendations of that theme were adopted in the design principles. Thereafter findings were used to reflect on and improve the design principles for formative assessment plans.

RESULTS

In this section, the outcomes of this study are presented through the subsequent steps that were taken to answer the research question.

1. Analysis and exploration: expert interviews
2. Design and construction: thematic analysis
3. Evaluation and reflection: group interviews with future users

Since the results will be presented in subsequent steps, they will reveal the creation as well as the evolution of the design principles for formative assessment plans so far, as advised by Bakker (2019).

Step 1: Analysis and Exploration: Expert Interviews

Table 4 shows the findings from the three expert interviews. The first expert group resulted in nine features that a formative assessment plan should have and the second and third expert group resulted in 11 features. The findings show some overlap between features that were mentioned in more than one group. The features goal orientation, alignment, giving insight in learning progress, leaving room for improvement, consciously and logical structured and involving competent teachers were mentioned in two or three expert groups. Because there were differences in descriptions and/or individual answers/aspects that were linked to these overlapping features, the features mentioned

in each group will be regarded and used as unique features at this point of the study. Consequently, the expert interviews resulted in a set of 31 unique features that a formative assessment plan should have according to the participants. The features are presented in **Table 4** in order of importance as ranked by the participants of the concerning groups.

Step 2: Design and Construction: Thematic Analysis

The 31 unique features that experts believe a formative assessment plan should have were used in a thematic analysis to derive themes for design principles.

The first column of **Table 5** shows the first themes that were made based on the features from the expert interviews. The second column represents the outcome of the critical review. And the third column shows the design principles that were formulated based on Van den Akker's structure for design principles (2013).

The second column of **Table 5** shows that two of the initial themes were not adopted. The active role of students, theme five, was not adopted because on closer inspection this was not considered as a feature specifically applicable for designing formative assessment plans. Theme eight, about prerequisites, was not adopted because it also did not represent a design principle for formative assessment plans rather conditions that should be in place before working with formative assessment plans.

The critical review also resulted in two themes being split up. On closer inspection theme alignment actually consisted of two themes: "Alignment of formative assessment activities inside and outside of the plan and with the rest of the curriculum" (3) and "Integration of formative assessment plans into curriculum/lesson plans" (4). And providing insight in learning processes was split up in two more specific themes: "Provide insight for teachers" (9) and "Provide insight for students" (10).

Overall, the thematic analysis resulted in 10 design principles for formative assessment plans.

Step 3: Evaluation and Reflection: Group Interviews

The draft version of the design principles was evaluated on their transparency, usability, completeness, or suitability by 23 teachers, four school leaders and one school policy maker during four group interviews (see **Table 3**). Through thematic analysis, three different types of recommendations to improve the design principles for formative assessment plans were found. These three themes were found in all interviews regardless of composition or context of the interviews.

The main points of improvement future users suggested for the 10 design principles were:

- Improve ambiguous writing
- Improve style and structure
- Improve content

Table 6 shows examples for each of these three themes.

According to the participants, improving ambiguous writing through more accessible concrete language enriched with practical examples and images would help make the design principles more suitable and usable for teachers in secondary education.

To improve style and structure participants suggested to put design principles in a chronological order and to present the design principles as an easy to use tool: roadmap, format, checklist, digital tool, or menu. Participants mainly linked these recommendations to enhancing transparency and usability of the design principles.

TABLE 4 | Results of expert interviews: 31 features, presented per expert group and ranked in order of importance by the participants.

	Expert group 1 (*N* = 5)	**Expert group 2 (*N* = 5)**	**Expert group 3 (*N* = 10)**
A formative assessment plan....			
	1. Is decision-driven	*10. Is goal oriented*	*21. Provides insight in learning*
	*2. Is goal oriented**	11. Stimulates an active role for students	*22. Is goal oriented*
	3. Provides insight in learning	*12. Leaves room for learning and improvement*	23. Clarifies success criteria for students
	4. Is logical structured	*13. Is aligned and evaluated with others*	24. Includes feedback, feedforward and feed-up
	5. Leaves room for improvement	*14. Provides insight in learning*	25. Has to take place in a safe and supportive learning environment
	6. Is effective	15. Prepares students for formative assessment	26. Teaches students to reflect on learning
	7. Is well-balanced	*16. Involves competent teachers*	*27. Involves competent teachers*
	Is aligned with other formative assessment plans	17. Includes hinge-points	28. Is widely applicable
	Is transparent	18. Defines next steps in learning	*29. Is consciously structured*
		19. Pays attention to different learning strategies	30. Provides tools and examples for formative assessment
		20. Is flexible	31. Leaves room for differentiation between students

**Italic text means that this feature was also mentioned in one of the other expert groups.*

TABLE 5 | Results of thematic analysis.

Themes of features	Critical review	Design principles (Draft to evaluate in group interviews) ***If you want to design formative assessment plans • then you are best advised to give these plans the following characteristics. . ..***
1. Consists of consciously chosen formative learning activities (4)*	Adopted	The plan has a logical and clear structure of formative assessment activities that build on from each other and are evenly spread. This means neither too much nor too little and sufficient variation in the formative assessment activities.
2. Is transparent (1)	Adopted	The plan must be transparent to all stakeholders. This means that those involved are aware and understand how formative assessment will be executed and why.
3. Is aligned (1)	Adopted and split up in two design principles	The plan consists of formative assessment activities that are aligned with each other, other formative assessment plans and the rest of the curriculum.
		The plan is integrated as much as possible into the curriculum/series of lessons.
4. Decision-driven (1)	Adopted	The plan consists of formative assessment activities that are consciously planned and chosen in the light of future decisions. It is clear in advance how the information provided by the formative learning activities will be used in making choices and decisions in (supporting) Students' learning.
5. Active role students (1)	Not adopted	
6. Flexible (3)	Adopted	The plan is flexible. Meaning that the plan creates room for moments of contingency. These moments of contingency can later be used to follow up on the information provided by formative assessment activities.
7. Leaves room for improvement (4)	Adopted	The plan leaves room for improvement and development of students. This means that after each formative activity there must be opportunity for all students to improve. Follow-up activities should be deliberately planned in order for students to use feedback. The formative assessment activity is therefore not only checking and concluding, but must be able to contribute to the next step in learning.
8. Prerequisites (8)	Not adopted	
9. Is goal-oriented (3)	Adopted	The plan is a set of consciously chosen formative assessment activities that are tied together by the same (or an overlap of) learning objectives.
10. Provides insight in learning (5)	Adopted and split up in two design principles	The plan provides insight into the learning processes of students at various times (how is a student doing, what development becomes visual, and what are the learning needs).
		The plan provides students with Insight into their own learning process at various times through feedforward of teachers or directly through formative assessment activities (how am I doing, what have I done so far, how can I continue?).

**The numbers in parentheses refer to the number of original features that were clustered under this category.*

TABLE 6 | Different types of recommendations with examples from the group interviews.

Recommendation	Examples of evaluation design principles	Examples of recommendations
Improve ambiguous writing	School B: "The document is still a bit vague, I miss a concrete explanation of activities" School C: "It wasn't always clear. We both read it in different ways"	School C: "Based on what you are trying to say here, also include an example of a teacher who is literally designing a lesson. So you can picture it a little better." School B: "You shouldn't use words as 'clear' and 'maybe.' I really think formulation can be more concrete"
Improve style and structure	School B: "If it is now (structured) in the right order, I think 10 is really much too late. Because 10 may move all the way forward, if you ask me, because that is about communication with others" School D: "but in terms of how (the set of design principles) is designed. If I have to work with this and then make a program (. . .) For me it's much better to work point by point	School A: just a very clear order of you going to do this first and then this and then this, which helps a lot. Chronology is important School B: "A lot of words, a more schematic display is preferable."
Improve content	School D "ten principles of which we say 'hey this one is almost the same as the others' Meaning that they overlap"	School A: "Especially merge (design principles). Off course automatically a few principles will disappear when you put things together. (. . .) Two and three can be put together"

Finally, participants thought that merging of overlapping texts and design principles would improve content and advance transparency, usability, and suitability of the design principles.

Final Adaptions Design Principles

To improve the design principles the following five actions were undertaken as a response to the outcomes of the group interviews. (1) Checking the design principles to see whether they could be formulated more concretely and to the point, (2) Checking the design principles for clear and consistent use of concepts, (3) Reviewing the design principles for overlap and repetition and merging if possible, (4) Reviewing the design principles to shorten sentences and texts where possible, and finally (5) Putting the design principles in a chronological order for designing formative assessment plans. The result of these actions is presented in **Table 7** and shows that part of design principle

TABLE 7 | Prototype design principles formative assessment plans.

Prototype design principles
1. Use a set of learning objectives and lesson plans as a starting point
2. Choose formative assessment activities that match the learning objectives that you are aiming for and the decisions you want to make
3. Plan formative assessment activities equally divided over time and in a way that they can build on from each other
4. Choose formative assessment activities that provide you with rich information about student learning and the necessary next steps in education and learning
5. Plan time, space and opportunity for students to improve their learning based on the outcome of formative assessment activities
6. Leave room for moments of contingency in formative assessment- and lesson plans
7. Align a formative assessment plan with other formative assessment activities that are taking place before, parallel or after this plan.
8. The plan must be transparent and feasible to all stakeholders

seven, from the draft version, is merged with draft version's principles 9 and 10 into design principle four. Draft design principles five and eight are combined in new principle two.

The suggestions made about a more schematic design and including examples and visuals in order to increase the usability will be considered at a later point in time when the design principles will be incorporated in a practical tool.

DISCUSSION

Formative assessment does not live up to the expectations when it is not carefully and coherently planned and constructively aligned with the rest of the curriculum (Wiliam and Thompson, 2007; Wiliam, 2013). In this study, a set of design principles for formative assessment plans was developed to support teachers in planning formative assessment activities coherently for the sake of well-informed formative decision-making.

Well-articulated design principles provide insight into the purpose and advised characteristics of an intervention accompanied by guidelines how to design this intervention, procedures, and conditions for implementation, all supported by empirical and theoretical arguments (Plomp and Nieveen, 2013; Bakker, 2019). It is important for future users that design principles provide this rich information to understand the value of the design principles together with when, why and how they work.

The design principles that are a result of the current study provide information about characteristics that formative assessment plans should have as well as procedures how to design these formative assessment plans. Often these characteristics and procedures can be recognized in literature regarding formative assessment activities that apparently often applies for formative assessment plans as well. In the next paragraph, the eight design principles will be used to give a preview on how these principles could be used in a design process.

The first four principles for formative assessment plans echo the importance of formative assessment activities to be aligned, coherent, and part of decision-driven data collection in order to be effective (Biggs, 1996; Wiliam, 2013; Furtak et al., 2016). As a result, *principle 1* advises teachers to use the learning objectives and existing lesson plans as starting point for their design of a formative assessment plan. Starting from learning objectives ensures that student learning is perceived in the light of learning processes toward general learning objectives instead of focusing on good or wrong answers (Coffey et al., 2011). Starting from existent lesson plans makes it easier for teachers to embed formative assessment activities in existing teaching processes and use existing learning activities as proof of learning to inform teaching (Earle, 2021). *Principle 2* recommends decision-driven data collection (Wiliam, 2013, 2014; Moss, 2020). Teachers determine in advance at which moments there is a need to make a decision about the next steps in teaching or learning with regards to the learning objectives. For example, decisions about adjusting lessons, how to differentiate, if students are ready for a new subject or what is the best way to support student learning at a given time. For these specific moments, teachers deliberately plan formative assessment activities that provide rich information about student learning on the defined learning objectives and helps inform the specified decisions (*Principles 3 and 4*). Deliberately planning these moments and formative activities linked to decisions and objectives ensures coherency and the possibility of formative activities to build on from each other.

Principle 5, 6, and 7 focus on how to make sure that the formative assessment plans leave room for improvements in teaching and learning. Formative assessment can only be effective if it results in a well-informed follow up and feedback can only become valuable for learners when they get opportunities to use it (Winstone and Boud, 2020). Concretely this means that after each activity that provides information about student learning teachers should plan time and possibilities for themselves and students to act upon this information (principles 5 and 6). Teachers must be able to adjust their lesson plans and students must be given the possibility to use feedback. Students should be provided with opportunities to improve their learning within the formative assessment plan. A recent study by Veugen et al. (2021) shows that teachers who use formative assessment mainly experience difficulties in making adjustments based on the outcomes of student learning. They do not always feel capable to make these adjustments. Therefore, a formative assessment plan should leave room for adjustments in teachers' instruction as well as the adjustments students want to make in their learning based on feedback they have received.

Principle 1 through 6 can be worked out in a timeline or added to a plan for a series of lessons. The final design principles, *principles 7 and 8*, focus on a final check of the formative assessment plan when everything is planned. Principle

7 challenges teachers to perceive their designed formative assessment plan in larger context while *principle 8* focuses on the check for transparency and feasibility of formative assessment plans to be useful and beneficial for all stakeholders.

Looking back at what defines the quality of design principles we can see that these eight design principles give information about procedures and characteristics that can help to design formative assessment plans. Nevertheless, thorough empirical support for these design principles lacks as the current study only consisted of theoretical evaluation with future users. Teachers have not had the chance to use these design principles in practice yet. Further investigation of the value and prescriptive validity of these principles in classroom practice is needed. A second limitation in this study is that, although this is an educational design study, future users were not part of all steps in the research process. Future users did prepare the thematic analysis in step two by clustering their answers during the expert interviews, however, the actual analysis in step two was conducted solely by researchers.

Bakker (2019) advises researchers, whenever they present design principles as outcomes of design research, to be explicit about the nature of the design principles. Are they values, criteria, predictions or advice (Bakker, 2019). At this moment these design principles contribute to existing literature on formative assessment by advising how to design formative assessment plans coherently, decision-driven and with successive and ongoing formative cycles. The design principles that are the outcome of the current design-study can be seen as an advice for teachers who want to design formative assessment plans. This might change into more prescriptive design principles in the future based on repeated cycles of design research. Nevertheless, the purpose will never be to formulate these design principles as strict guidelines (Havnes et al., 2012). The main goal is that these design principles can support teachers now and in the future to design decision-driven formative assessment that informs their teaching and improves learning.

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 Ethische Commissie Onderzoek HAN university of applied sciences. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JS was responsible for the research work and preparing the manuscript. TS-M, CV, and DJ-T were involved in publication planning, providing advice on directions for the research and manuscript, reading of drafts and giving suggestions for changes and feedback on the final manuscript.

FUNDING

This work was supported by the Taskforce for Applied Research SIA, or Regieorgaan SIA (Dossier no. RAAK.PRO03.057).

ACKNOWLEDGMENTS

We wish to thank all who contributed to completing this study and the set of design principles. Special considerations for the schools who are our committed partners in learning about and designing formative assessment plans.

REFERENCES

Antoniou, P., and James, M. (2014). Exploring formative assessment in Primary School Classrooms: developing a Framework of Actions and Strategies. *Educ. Assess. Eval. Account.* 26, 153–176. doi: 10.1007/s11092-013-918

Bakker, A. (2019). "Design principles in design research: a commentary," in *Unterrichtsentwicklung Macht Schule*, ed. M. Peters (Wiesbaden: Springer), 1–7. doi: 10.1007/978-3-658-20487-7

Bearman, M., and Dawson, P. (2013). Qualitative synthesis and systematic review in health professions education. *Med. Educ.* 47, 252–260. doi: 10.1111/medu.12092

Biggs, J. (1996). Enhancing teaching through constructive alignment. *High Educ.* 32, 347–364. doi: 10.1007/bf00138871

Black, P., and Wiliam, D. (2009). Developing the theory of formative assessment. *Educ. Assess. Eval. Account.* 21, 5–31. doi: 10.1007/s11092-008-9068-5

Black, P., and Wiliam, D. (2010). Inside the Black Box: raising Standards Through Classroom Assessment. *Phi Delta Kappan* 92, 81–90. doi: 10.1177/003172171009200119

Briggs, D. C., Ruiz-Primo, M. A., Furtak, E., and Shepard, L. (2012). Meta-Analytic Methodology and Inferences About the Efficacy of Formative Assessment. *Educ. Measure.* 31, 13–17. doi: 10.1111/j.1745-3992.2012.00251.x

Christoforidou, M., Kyriakides, L., Antoniou, P., and Creemers, B. P. M. (2014). Searching for stages of teacher's skills in assessment. *Stud. Educ. Eval.* 40, 1–11. doi: 10.1016/j.stueduc.2013.11.006

Coffey, J. E., Hammer, D., Levin, D. M., and Grant, T. (2011). The missing disciplinary substance of formative assessment. *J. Res. Sci. Teach.* 48, 1109–1136. doi: 10.1002/TEA.20440/FORMAT/PDF

Deneen, C. C., Fulmer, G. W., Brown, G. T. L., Tan, K., Leong, W. S., and Tay, H. Y. (2019). Value, practice and proficiency: teachers' complex relationship with assessment for learning. *Teach. Teacher Educ.* 80, 39–47. doi: 10.1016/j.tate.2018.12.022

Earle, S. (2021). Formative Decision-Making in Response to Primary Science Classroom Assessment: what to do Next? *Front. Educ.* 5:584200. doi: 10.3389/FEDUC.2020.584200/FULL

Fjermestad, J., and Hiltz, S. R. (2000). 'Group support systems: a descriptive evaluation of case and field studies'. *J. Manage. Information Syst.* 17, 115–159. doi: 10.1080/07421222.2000.11045657

Furtak, E. M., Kiemer, K., Circi, R. K., Swanson, R., de León, V., Morrison, D., et al. (2016). 'Teachers' formative assessment abilities and their relationship to student learning: findings from a four-year intervention study'. *Instructional Sci.* 44, 267–291. doi: 10.1007/s11251-016-9371-3

Furtak, E. M., and Ruiz-Primo, M. A. (2008). Making students' thinking explicit in writing and discussion: an analysis of formative assessment prompts. *Sci. Educ.* 92, 799–824. doi: 10.1002/sce.20270

Gu, P. Y. (2021). An Argument-Based Framework for Validating Formative Assessment in the Classroom. *Front. Educ.* 6:605999. doi: 10.3389/FEDUC.2021.605999

Gulikers, J., Gulikers, J. T. M., Biemans, H. J. A., Wesselink, R., and van der Wel, M. (2013). Aligning formative and summative assessments: a collaborative action research challenging teacher conceptions. *Stud. Educ. Eval.* 39, 116–124. doi: 10.1016/j.stueduc.2013.03.001

Havnes, A., Smith, K., Dysthe, O., and Ludvigsen, K. (2012). Formative assessment and feedback: making learning visible. *Stud.Educ. Eval.* 38, 21–27. doi: 10.1016/j.stueduc.2012.04.001

Heitink, M. C., Heitink, M. C., van der Kleij, F., Veldkamp, B. P., Schildkamp, K., and Kippers, W. B. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educ. Res. Rev.* 17, 50–62. doi: 10.1016/j.edurev.2015.12.002

Kingston, N., and Nash, B. (2011). Formative assessment: a meta-analysis and a call for research. *Educ. Measure.* 30, 28–37. doi: 10.1111/j.1745-3992.2011.00220.x

McKenney, S., and Reeves, T. C. (2019). *Conducting Educational Design Research.* 2nd Edn. London: Routledge.

Moss, C. M. (2020). "Role of Educational Leadership in Confronting Classroom Assessment Inequities, Biased Practices, and a Pedagogy of Poverty," in *Handbook on Promoting Social Justice in Education*, ed. R. Papa (Cham: Springer), 863–887. doi: 10.1007/978-3-030-14625-2_147

Nowell, L. S., Norris, J. M., White, D. E., and Moules, N. J. (2017). Thematic analysis: striving to meet the trustworthiness criteria. *Int. J. Qual. Methods* 16, 1–13. doi: 10.1177/1609406917733847

Offerdahl, E. G., McConnell, M., and Boyer, J. (2018). Can I have your Recipe? Using a Fidelity of Implementation (FOI) Framework to Identify the Key Ingredients of Formative Assessment for Learning. *CBE Life Sci. Educ.* 17:es16. doi: 10.1187/cbe.18-02-0029

Plomp, T., and Nieveen, N. (2013). *Educational Design Research*. Netherlands: Netherlands Institute for Curriculum Development (SLO). doi: 10.1007/978-1-4614-3185-5_11

Pyrko, I., Eden, C., and Howick, S. (2019). Knowledge Acquisition Using Group Support Systems. *Group Decision Negotiation* 28, 233–253. doi: 10.1007/s10726-019-09614-9

Ruiz-Primo, M. A., and Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *J. Res. Sci. Teach.* 44, 57–84. doi: 10.1002/tea.20163

Schildkamp, K., and Kuiper, W. (2010). Data-informed curriculum reform: which data, what purposes, and promoting and hindering factors. *Teach. Teacher Educ.* 26, 482–496. doi: 10.1016/j.tate.2009.06.007

Van den Akker, J. (2013). "Curricular development research as a specimen of educational design research," in *Educational Design Research*, eds T. Plomp and N. Nieveen (Netherlands: Netherlands Institute for Curriculum Development (SLO)), 52–71.

Van Den Berg, M. (2018). *Classroom Formative Assessment A Quest for a Practice that Enhances Students*. Netherland: University of Groningen.

Van der Kleij, F. M., Vermelulen, J., Schildkamp, K., and Egen, T. (2015). Integrating data-based decision making, Assessment for Learning and diagnostic testing in formative assessment. *Assess. Educ. Princ. Policy Pract.* 22, 324–343. doi: 10.1080/0969594X.2014.999024

Veugen, M. J., Gulikers, J. T. M., and Den Brok, P. (2021). We agree on what we see: teacher and student perceptions of formative assessment practice. *Stud. Educ. Eval.* 70:101027. doi: 10.1016/j.stueduc.2021.101027

Wiliam, D. (2011). What is assessment for learning?. *Stud. Educ. Eval.* 37, 3–14. doi: 10.1016/j.stueduc.2011.03.001

Wiliam, D. (2013). Assessment: the Bridge between Teaching and Learning. *Voices Mid.* 21, 15–20.

Wiliam, D. (2014). *Formative Assessment and Contingency in the Regulation of Learning Processes*. Available online at: http://www.dylanwiliam.org/Dylan_Wiliams_website/Papers_files/Formative assessment and contingency intheregulation of learning processes (AERA2014).docx (accessed January 15, 2020).

Wiliam, D., and Thompson, M. (2007). "Integrating Assessment with Learning: What Will It Take to Make It Work?," in *The Future of Assessment*, ed. C. A. Dwyer (Mahwah, NJ: Erlbaum), 53–82.

Winstone, N. E., and Boud, D. (2020). The need to disentangle assessment and feedback in higher education. *Stud. High. Educ.* 47, 656–667. doi: 10.1080/03075079.2020.1779687

Wylie, E. C., and Lyon, C. J. (2015). The fidelity of formative assessment implementation: issues of breadth and quality. *Assess. Educ.Princ. Policy Pract.* 22, 140–160. doi: 10.1080/0969594X.2014.990416

Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., and Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assess. Educ. Princ. Policy Pract.* 28, 228–260. doi: 10.1080/0969594X.2021.1884042

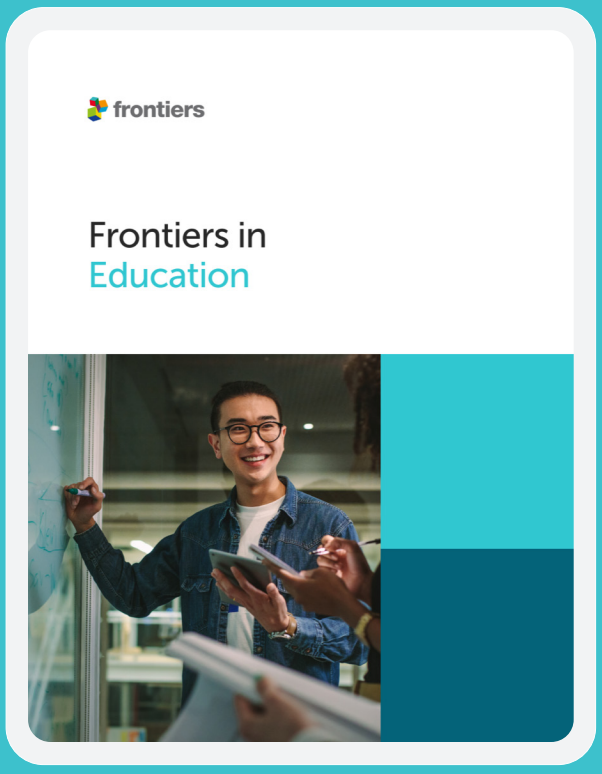
Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Frontiers in Education

Explores education and its importance for individuals and societyA 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

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact