Reasoning in classroom dilemma situations: how pre-service teachers judge performance assessment

Ulrike Greiner*, Michaela Katstaller and Theresa Oitner

Department of Educational Science, Paris Lodron University, Salzburg, Austria

This study examines pre-service teachers' reasoning structures based on their beliefs in the context of school performance assessment. We used reflective writing to investigate pre-service teachers' judgment and reasoning regarding different functions of performance assessment. Forty-five undergraduate pre-service teachers participated in our study. Using a mixed-method approach, we conducted categorial and reconstructive text analyses as well as exploratory statistical analyses to describe the participants' reasoning structures. Such cognitive structures comprise potential solutions to the performance assessment dilemmas that teachers face in their everyday teaching practice. We found varying distributions of and relationships between (individual-, objective-, social-, and ability-related) reference norms (neutral, student-, and teacher-centered) reference perspectives as well as (causal-analytic, normative, descriptive, and effect-oriented) modes of argumentation. Our discussions related to future research activities on teachers' reasoning structures in the classroom.

1 Introduction

Teaching is currently considered a dynamic, epistemic, ethical, and social practice based on judgments that are shaped by teachers' cognitions and beliefs (Loughran et al., 2016). In the field of research on teacher-knowledge, there is strong evidence of the important impact of teachers' cognitions and beliefs on judgment processes in classroom settings (e.g., Shavelson and Stern, 1981; Neuweg, 2014). Recently, classroom settings and related educational requirements have changed in the sense that educational challenges that teachers face have become more ill-structured, complex, and sometimes contradictory (Schuck et al., 2018). Consequently, teachers – especially pre-service teachers – must acquire skills that allow them to handle conflicting claims and uncertainty in everyday teaching, to deal with educational dilemmas, and to make decisions regarding complex and diffuse educational issues. Given that such decisions are assumed to affect behavior, Shulman (1987) states that 'judgment, rather than behavior, is the essence of teaching' (as cited in Neuweg, 2014, p. 583). Judgment is shaped by underlying beliefs; therefore, teachers' beliefs and related reasoning structures play a significant role in their decision-making (e.g., Penso and Shoham, 2003).

This is particularly true for socially and individually important and often controversial performance assessment activities in the classroom. Assessing students' performance

represents a core activity before, during, and after teaching that occurs daily or several times a day inside and outside the classroom (Terhart, 2014). Assessment-related decision-making is influenced by various and often opposing factors such as curriculum, educational, or evaluation standards as well as individual and group-based responsibilities, assessment practices or cultures, institutional requirements, or individual students' or parents' needs (Zhang and Burry-Stock, 2003; Pope et al., 2009; Jones and Lawson, 2018). Assessment and certificates, such as school-leaving certificates, have been of considerable relevance lately, so they can have a strong impact on those individuals and their careers who carry out performance assessment activities. Therefore, teachers' decisions concerning assessment are intensively debated among parents. It is hardly surprising that especially pre-service teachers or early-career teachers have to contend with significant problems, concerns, and insecurities about performance assessment and related decision-making (Beziat and Coleman, 2015). Therefore, the present study aims to examine pre-service teachers' beliefs regarding performance assessment by evaluating their decision-making processes based on reasoning structures.

2 Theoretical framework

Research on teachers' assessment activities has not only focused on grading practices, standardized testing, coursework, assessment standards, or alternative forms of assessments but also on teachers' beliefs about assessments (Campbell, 2013). There is significant evidence for the reciprocal relationship between beliefs, judgments, and reasoning structures in the teaching profession (Bendixen et al., 1994; Loibl et al., 2020). Based on these findings, the theoretical background of this study focuses on how pre-service teachers' beliefs about assessment shape their judgment. In particular, we focus on reasoning structures that show pre-service teachers' modes of argumentation with regard to decision-making when it comes to arguing about performance assessment.

2.1 Pre-service teachers' beliefs

Pre-service teachers' beliefs are theoretically anchored in various constructs as part of their professional competence. According to Reusser and Pauli (2014), beliefs are attitudes with both affective and normative characteristics. Teachers' beliefs are personal attitudes unlike norms and come into play in the analysis of the argumentative structures of communication. Although norms, especially reference norms, i.e., central substantive categories of evaluations, express distant societal expectations, they are still used individually and embedded in arguments to give shape to one's own beliefs (Fives and Gill, 2014).

Beliefs with strong cognitive links to actions are particularly important in teachers' socially relevant fields of action, such as performance assessment. Beliefs about contents, people, and contexts are related to each other in academic performance assessment. Academic assessment has a considerable impact on people and cannot be separated from the specific context in which it is conducted; therefore, research into performance assessment must consider the contexts of schools and teachers.

In our study, we focus on the content-related dimensions of pre-service teachers' beliefs about performance assessment related to the subject matter to be taught but also to social knowledge about people (preferably students) in the contexts of school and society. In particular, we consider 'epistemic beliefs' (Franco et al., 2012) which influence 'the way in which individuals look at the world … in order to gain knowledge' and which describe an individual's beliefs about knowledge itself and theories of knowledge (Maggioni and Parkinson, 2008, p. 447). Hofer and Pintrich (1997, p. 112) assumed that epistemic beliefs 'are not organized into stages or levels' as suggested in Perry's (1981) model and should therefore be treated instead as subjective theories with a normative connotation as they reveal traits of personal attitudes and philosophies. Pre-service teachers' epistemic beliefs are evident in their selection of teaching methods and materials as well as their behavior in teaching situations (Aguirre and Speer, 2000). Their beliefs not only affect their decisions regarding what to teach and how to act in the classroom but also how to assess students' learning progress and performance (Ioannou-Georgiou and Pavlou, 2003).

2.2 Performance assessment

Performance assessment is a goal-based activity designed 'to measure a skill or ability' (Frey and Schmitt, 2007, p. 416). Within classroom settings, performance assessment means not only to measure skills or abilities but also to use measurements to interpret performance in relation to standards as well as to issue grades to students. This process is often linked, for example, to questions of reliability and validity as well as to the tension between the quality of the assessment tool and the fairness of the tester (Halkes, 1981; American Educational Research Association et al., 2014; Reh and Ricken, 2018). In combination, measuring, interpreting, and grading represent a highly complex process (DeLuca et al., 2018). Such high levels of complexity in classroom settings increase even more when there is no explicit definition of what teachers should measure and how these measurements should be interpreted (Neuweg, 2019). Hence, it seems very likely that when a teacher's (rational) knowledge and skills are limited (less rational) beliefs have a major impact on decision-making related to performance assessment (Xu and Brown, 2016).

In terms of the pre-service teachers, who took part in this study, it can be said that the essential basics of assessment skills regarding examination performance is part of their curriculum. Thus, while they acquire knowledge about the theoretical foundations in this respect, the students still lack practical training within their program.

2.3 Dimensions of beliefs about performance assessment

Research findings on (pre-service) teachers' beliefs about performance assessment in classroom settings indicate that it is important to differentiate between the various dimensions of such beliefs (e.g., Barnes et al., 2015; Schmidinger et al., 2015). These dimensions concern (1) the purpose of assessment (i.e., promotion or selection), (2) the positions or roles of students and teachers during assessment, (3) measurement standards like objectivity, reliability, and validity as well as related problems, (4) the relationship between

performance recording, performance assessment, and different modes of assessment (e.g., oral or written), and (5) the impact of assessment on students' learning and achievement.

The debate within teacher education about the relationship between fairness, validity and reliability in formative and summative assessment is persistent until today (Harlen, 2005; McMillan, 2011). Measurement theory agrees upon the fact that fairness is distinct but related to validity and reliability (Stobart, 2006). Both fairness and validity cannot be determined dichotomously because it is a matter of degree, just reliability (Cole and Zieky, 2001). According to Camilli (2006), fairness relates to "factors beyond the scope of the test" (p. 225). Most research on fairness in educational assessment refers to large-scale assessment. However, recent changes in the educational landscape allow to investigate fairness as a quality of classroom assessment.

In performance assessment processes, teachers must straddle these different dimensions which often leads to a 'dilemma' defined as 'a situation that makes problems, often one in which you have to make a very difficult choice between things of equal importance' (Oxford University Press, 2023). In teacher education at our university, preservice teachers reflect on the principles of fairness that precede an assessment (e.g., access to learning documents), determinants during the assessment (e.g., design) and its consequences (e.g., the interpretation of results) (Baniasadi et al., 2023).

2.4 Performance-assessment dilemmas

Generally speaking, the process of teaching can be characterized by conflictual situations and contradictory demands that lead to a range of dilemmas. Wegner et al. (2014, p. 46) postulated that understanding teaching means recognizing the 'dilemmatic nature of teaching' and considering five different types of dilemmas that can arise in educational contexts at any time: the dilemma of self-regulation, the dilemma of didactic structure, the heterogeneity dilemma, the dilemma of professional relationship with learners, and the assessment dilemma. According to Suurtamm and Koch (2014), the performance assessment dilemma is related to conceptual, pedagogical, cultural, and political dilemmas. How the different dilemmas are justified and acted out in the classroom influences the quality of teaching and assessment. For example, one performance assessment dilemma is related to the tension between different reference norms of performance in the assessment itself. During the assessment process, teachers must decide how to balance criteria-based norms (e.g., the achievement of goals from curricula), social norms (e.g., the performance of other students) and individual norms (e.g., individual performance history). Tensions between the social allocation function and individual support as the primary functions of schools represent another dilemma. There is currently a trend to shift from 'a view of assessment as an event that objectively measures the acquisition of knowledge toward a view of assessment as a social practice that provides continual information to support student learning' (Suurtamm and Koch, 2014, p. 264). Performance assessment in schools has long been considered a controversial issue because schools have primarily been seen as institutions for allocation on the one hand and as institutions supporting individual development on the other hand (Schmidinger et al., 2015; Breidenstein, 2018). A further dilemma arises regarding how to handle the measurement

standards of objectivity, reliability, and validity as well as the dynamics of formative assessment in daily classroom settings. There is a tendency to 'include students in developing and applying assessment criteria' (Suurtamm and Koch, 2014, p. 265), which requires teachers to balance generally and individually valid standards of performance assessment. In addition, the variability of national and international standardization of performance appraisal and the professional judgment of teachers can lead to problems because predetermined curriculum goals or national assessment outcomes can vary according to teachers' varied understandings of how to assess the diverse learners in their classrooms (Fives et al., 2017). Finally, tensions and related dilemmas arise between different types of assessment, such as between written and oral assessment: For instance, oral assessment can sometimes cause a dilemma because its inclusion in students' overall grading is established by law to encourage teachers to continuously observe and assess students' learning processes (Amrhein-Kreml et al., 2008; Neuweg, 2019). However, standardized exams are typically conducted in written form to guarantee objectivity, reliability, and validity. Thus, the combination of oral and written types of assessments puts pressure on teachers' assessment processes. Many of these examples of tensions and dilemmas were considered when choosing an input for the reflective writing task in our study; moreover, these dilemmas provided the core source of educational judgment and reasoning.

2.5 Educational judgment, educational reasoning, and reasoning structures

To handle educational dilemmas successfully, teachers must constantly assess and judge educational situations; consequently, teaching as well as performance assessment becomes a complex judgment process that entails 'comprehension, reasoning, transformation, evaluation and reflection' (Penso and Shoham, 2003, p. 315). Educational judgment depends on personal characteristics (e.g., knowledge, beliefs, or attitudes), situational properties (e.g., goals or time pressure), behaviors of teachers (e.g., verbalizations), thinking processes (e.g., perceiving, interpreting, or decision-making) as well as other impact factors (e.g., Loibl et al., 2020).

Educational judgments are based on cognitive processes that integrate beliefs 'in a plausible narrative that allows understanding of a situation' and socially situated processes that are 'shaped by the exchanges among actors and by the systemic features of an organizational and cultural context' (Allal, 2013, p. 23).

Judgments are based on 'reasoning' which represents a process in which arguments are put forward and evaluated (Shaw, 1996). Pre-service teachers' educational reasoning has become increasingly important, especially in making decisions concerning difficult classroom situations, complex pedagogical problems, and conflicts with students, parents, school leaders, or other stakeholders (Guerriero, 2017). Recently, controversial issues of modern science that involve social, political, economic, ethical, and pedagogical considerations appear in educational contexts. Such conflictual situations require teachers to critically elaborate and evaluate potential solutions because they cannot simply be resolved by applying 'cause and effect reasoning' (Eggert et al., 2012, p. 3). Such situations also require informal reasoning which deals with ill-defined problems in response to complex issues and solutions that make conflicting reasons meaningful

(Wu and Tsai, 2011; Fang et al., 2019). Unlike formal reasoning, which uses logical rules to address a problem, informal reasoning in assessment situations requires pre-service teachers to construct and evaluate their arguments on ill-structured problems (Sadler, 2004).

Although teachers' judgment and related reasoning processes represent crucial components of performance assessment, little research has been done on this subject (Wyatt-Smith et al., 2010; Spooner-Lane et al., 2022). According to Toulmin (2003), reasoning structures consist of topics and claims, grounds or evidence provided through different modes of argumentation, and perspectives. Research has been carried out on teachers' pre-actional, actional, and post-actional reasoning structures in the context of assessment and diagnostic competence (Klug et al., 2013). However, such research has not yet considered cognitive reasoning structures and processes in detail, but rather has focused on products of reasoning like tests and self-assessments. Moreover, a qualitative study of teacher assessment by Remesal (2011) investigates the effects of assessments on teaching, learning, the accountability of teachers and school, and the measurement of achievement. Such conceptions, though, tend to be related to stable beliefs and not to dynamic reasoning structures.

3 Purpose of the study

The present study aims to analyze reasoning structures related to pre-service teachers' beliefs about performance assessment. In general, our study focuses on how pre-service teachers attempt to establish a 'pedagogical equilibrium' by balancing different influencing and often conflicting factors (Loughran, 2019, p. 527). We focused on pre-service teachers because they experience educational misconceptions during their studies which shape their beliefs about performance assessment in the classroom (Menz et al., 2021). Qualitative or mixed-method studies that address the cognitive aspects of pre-service teachers' assessment skills and activities, like our study does, only focus on assessment activities often neglecting the cognitive process perspective (Ogan-Bekiroglu and Suzuk, 2014). Based on the reported backgrounds, the following questions arise for this study: (1) Which reasoning structures do pre-service teachers use in assessment dilemma situations?; (2) How are the elements of reasoning structures distributed quantitatively, and which correlations of sub-categories shaping the judgments can be discovered?; and (3) Which implicit or latent structures of meaning, which shape social practice by dealing with the tensions of performance assessment dilemmas, become visible in their reasoning?

4 Methods

4.1 Writing task

Through a reflective writing task, we extracted pre-service teachers' reasoning structures, namely, reference norms, reference perspectives, and modes of argumentation, in relation to performance assessment dilemmas. Reflective writing has become a highly valued approach in teacher education because it enables participants to articulate their views and beliefs on educational issues (e.g., Cohen-Sayag and Fischl, 2012). Furthermore, it serves as an effective instrument to uncover modes of argumentation, which are otherwise

difficult to measure (e.g., Shavelson et al., 2019). The problems in terms of reliability, validity and fairness, which are addressed in the following writing task, are well known in daily performance-assessment-practices in Austrian schools. The textual input for the writing task is taken from a scientist's speech addressed to teachers:

> Imagine that you, as a pre-service teacher during your internship, are asked to state your opinion on the following statement of the scientist Rainer Dollase. Please explain which response you would give to your students and support your arguments using approximately 150 words.
>
> *In Finland, grading almost exclusively happens by assessing written performances, e.g., by making use of traditional in-class examinations. For many years, I have been pointing out that the assessment of oral contributions disadvantages quiet students, but above all, it leads to the development of an interesting 'chattering culture.' For example, one student approached me, saying, 'I would've never passed my school leaving exam in Math, if oral class participation had not contributed to my grade.' When I asked them how this was possible, the same student responded: 'I always asked questions like: How did you get from line 7 to the result in line 8?' The teacher thought that I was genuinely interested in Math and, because of that, he compensated my negative grade in written exams with positive grades in oral performances.' Class participation surely is nice and relieving; however, an individual incentive to make a greater effort can only be achieved by assessing oral and written performances individually, for example, by conducting individual oral examinations or alternatively by written tests* (Dollase, 2004).

4.2 Participants

Data were collected during the winter term of 2018/2019 at the School of Education at the Paris Lodron University of Salzburg. A sample of 48 undergraduate pre-service teachers in two courses on research methods and individual learning support participated in the study. No benefits for study participation were given. Data collection was administered in one session and lasted approximately 45 min. Subjects participated voluntarily and were assured that their anonymity would be protected and that the researchers would comply with data protection regulations. In total, 45 participants completed the writing task and were included in our data. The age of the participants ranged from 19 to 37 years ($M = 24.07$, $SD = 4.85$); 64% were female and 36% were male.

4.3 Data analysis

For our study, we applied a three-pronged approach together with a mixed method strategy on data analyses. By applying a categorical content analysis, we first identified which reference norms, reference perspectives, and modes of argumentation pre-service teachers use.

Second, we focused on combinations of categories to identify correlational patterns in reasoning structures (see Table 1). Third, during the summer term of 2020, we conducted a secondary analysis (Medjedovic, 2014), namely a qualitative analysis to reconstruct latent structures.

TABLE 1 Frequencies of codes in triple combinations of reasoning structures.

				Modes of argumentation				
				Causal-analytic	Normative	Descriptive	Effect-oriented	Sum
Reference norm	Individual reference norm	Reference perspective	Neutral	0	8	4	1	13
			Student	1	0	2	1	4
			Teacher	6	9	11	6	32
	Objective reference norm		Neutral	0	1	4	1	6
			Student	0	0	1	0	1
			Teacher	0	9	12	0	21
	Social reference norm		Neutral	1	1	0	0	2
			Student	0	0	1	0	1
			Teacher	1	1	2	2	6
	Ability-oriented reference norm		Neutral	0	0	2	6	8
			Student	0	0	2	0	2
			Teacher	2	2	3	4	11
Sum				11	31	44	21	107

Across all stages of the data analyses, we used a mixed method strategy focusing on an 'exploratory-sequential approach' (Edmonds and Kennedy, 2013, p. 167). To do this, we combined qualitative data with exploratory and non-parametric statistical procedures with a focus on categorial data taken from small samples (e.g., Bortz and Lienert, 2003). We quantified the qualitative results and tested for distribution patterns of single variables (using one sample distribution tests), for relationships between variables (using chi-squared tests), and for implicit resp. latent structures (using dummy-coding and exploratory factor analysis).

The exploratory sequential technique offers the benefit of enhancing quantitative results through the incorporation of qualitative data. Thus, quantitative data analyses explain results in succession and allow for greater versatility in discovering novel ideas in a qualitative approach (Gogo and Musondo, 2022).

The participants' written responses were collected and analyzed using content analysis (Mukherjee et al., 2018). Content analysis is a social-scientific method that examines textual material embedded in its original context and refers to procedures for assessing the relative extent to which certain themes, attitudes, motifs, and beliefs permeate certain documents or messages (e.g., Mayring and Gläser-Zikuda, 2008; Gläser and Laudel, 2010; Mayring, 2015). Our categorial content analysis of the participants' texts focuses on the underlying patterns of argumentation with a view to developing not just a broader understanding of professional teaching but also of the argumentative process (Konstantinidou and Macagno, 2012). The data were analyzed in a mixed coding process, applying both a deductive and an inductive approach (e.g., Gholami and Husu, 2010).

4.4 Coding process

The collected texts were categorized in MAXQDA by two individual coders (Rädiker and Kuckartz, 2019), both of whom were trained and experienced research assistants. Coding and analysis were conducted by the first and fourth author of the study. Before commencing the categorization process, both coders had to first scan the texts without assigning them to any category. Following the theoretical framework, we used the presented model of categories (see Table 1). After the preliminary scan, the coders discovered that the texts must be divided into units of meaning. In an abductive procedure, the research group (consisting of the first, third and fourth author) agreed to assign units of meaning to the categories' reference norms, reference perspectives, and modes of argumentation. Each unit was examined and assigned a code for each category (complete coding). Virtually, every text consists of more than one unit of meaning, with an average of two to three units per text. Overall, we had 107 units in 45 texts.

Each coder worked individually on the participants' written texts. During the initial coding phase, the coders were in contact with the first author to ensure that the categories and subcategories were applied consistently. Each coder conducted 321 codings (i.e., 107 units on three subcategories). After categorizing the texts independently, the coders merged their codings in MAXQDA. Next, the intercoder-reliability was calculated to determine the extent to which the two coders made the same codings after independently evaluating the texts (Lombard et al., 2002). The Cohen's kappa (k), which indicates the reliability of the coding, is 0.86 (the mean of the 45 texts collected). In 24 texts, the coders identified discrepancies concerning the assignment of categories; thus, they compared the nonconforming codings and were able to reach a resolution with the help of the defined category system. The full sample was then used to analyze the results.

When creating the category system through an abductive procedure, our main goal was to find out whether and how the pre-service teachers' beliefs were related to the reflective structures of their texts. After the analysis had taken place and been monitored by the first author, the fourth author reviewed the coding process once more.

4.5 Secondary analysis of latent structures

The secondary analysis is devoted to our third research question on whether implicit structures of meaning become visible in the

written texts. From the perspective of reconstructive qualitative research, teachers' beliefs and practice dealing with assessment dilemmas are shaped by implicit knowledge that is collectively shared. To learn more about these implicit or latent structures underlying teachers' decision-making, we carried out a reconstructive analysis of the collected texts by employing the documentary method (Reischl and Plotz, 2020). Based on the methodological background of Mannheim's sociology of knowledge (Bohnsack, 2017), we delved deeper into the implicit structures of meaning to explicate an implicit knowledge on how such dilemma situations can be dealt with or nevertheless be resolved.

For this analysis, we re-read all the gathered texts. We then pre-selected 10 of the 45 texts in the way of sorting (Medjedovic, 2014) by taking a sub-sample of critical cases (Flick, 2013), thus limiting the analysis so as to learn more about pre-service teachers' strategies for assessment dilemmas. Our research question as stated above acted as the second important criterion for text selection. The 10 selected texts not only directly addressed the main topics of the performance assessment dilemma, but also offered different aspects of problem-solving against the background of a still unexplained contextual knowledge of the social field of teaching. These 10 texts were subjected to the analytical steps of the documentary method (such as formulating, reflective interpreting, and first type formation) (Schäffer, 2020).

5 Results

5.1 Qualitative results of the content analysis on reasoning structures

In this section, we describe the results of the content analysis of the pre-service teachers' reasoning structures regarding performance assessment dilemmas related to reference norms, reference perspectives, and modes of argumentation. We present descriptions of different categories as well as illustrative examples of text passages that illustrate these descriptions.

5.1.1 Results on reference norms

5.1.1.1 Individual reference norm (equates ipsative-referencing)

Within the analyzed texts, the individual reference norm appeared most frequently (49 times; see the corresponding row sum in Table 1). The students' individual needs, capabilities, and developments are the most prominent themes that emerge in relation to the selection of an appropriate form of assessment and evaluation:

> We live in an era of individualization, which primarily means to pave the best possible way for everybody, to develop oneself. Therefore, quiet students should be offered the possibility to improve their grades by written tests whereas students who are keen to debate by oral tasks. (ID 010)

We also found indications that individual reference norms are reflected in the context of oral or written forms of assessment. Oral performance assessment is often associated with ideas such as individual learning experience, students with special needs,

compensation for weaknesses, individuality, and opportunities for personal interaction between students and teachers. However, written performance is more often linked to concepts such as standardization and objectivity. Although possible tensions between different modes of assessment (oral, written) and measurement standards like objectivity and validity exist, the writer assumes the application of both oral and written forms of examination without setting on one assessment option. In this case, both types of assessment deliver information on student performance related to an individual reference standard:

> That written assessment also has advantages, is out of question: 'What's agreed is agreed!' For the teacher, this is a convenient way, besides the effort of reading through and correcting, to form an opinion about a person. (ID 011)

Another aspect of assessment related to an individual reference norm concerns the method of examination and the active decision-making role of students within the assessment process. Considering each student's perspective, the standard of assessment should be unique to the student, who, in some circumstances, should be allowed to choose the method of examination:

> Summed up, I would say that it makes sense to choose the method of examination with attention to the student. Depending on the strengths and weaknesses [of the student], an appropriate method of examination makes sense. Maybe it could also be of importance to let the students choose the method of examination themselves. (ID 012)

5.1.1.2 Objective reference norm (equates criteria-referencing)

In the context of school performance assessment, the objective reference norm provides the subject-specific criteria against which student performance is to be measured. The reference to the school subject and the curricula often appears to be the most stable factor, independent of human individuals and groups, but certainly demanding in terms of the diverse content and skills required. An objective reference norm appeared less often (28 times). This is characterized by the consideration of the respective curriculum, subject, and/or special goal-related features of the school. The choice of the objective reference norm as a category of reasoning structures for performance assessment was also found to be related to the emphasis on the diversity of perspectives on performance or the difference between various norms:

> To answer a question to the students of that kind it is necessary to view the question from various perspectives. It needs to be mentioned that within the current curriculum students should acquire different competences in the subjects. (ID 013)

5.1.1.3 Ability reference norm

The ability reference norm appeared 21 times and concerns skill-and-competencies-related aspects of what students can do with learned subject matter contents. It was linked to different topics in pre-service teachers' texts. The strongest link is that of the students' study and their professional career after school. Less distinctive is the connection to the abilities resp. competencies prescribed by the

curriculum. The assessment is oriented toward the question of whether the method of performance assessment concerning knowledge or ability is relevant for the future, especially for each student's prospective career and professional life:

> The answer to this question consists of several parts. On the one hand, the current curriculum strongly focuses on competences, including the social competence as a main factor. For the actual prospective working life, the social competence is even more important than it is at school, which is why I actually welcome the idea of an oral influence on the grade. (ID 014)

5.1.1.4 Social reference norm

In view of the increased importance of individualized feedback in the sense of formative assessment, but also of self-regulated learning, the social norm seems to be losing importance (especially for the next generation of teachers?), but on the other hand it is still valid in terms of teachers' practices that implicitly focus on the comparison of different performance groups (Hofmann et al., 2016).

Compared to the other reference norms, the social reference norm which relates performance assessment to social comparisons with other students appears only rarely (nine times). The following text sequence illustrate the consideration of the social reference norm:

> Discussions help students to understand others' perspectives, to rethink their views, to defend [their] points of view, to make compromises and much more. The exchange with other students can help students to understand tasks that they have not understood before and therefore achieve better results regarding their performance. (ID 015)

5.1.2 Results on reference perspectives

The reference perspective category shows from which individual-or situation-related perspectives pre-service teachers formulate their arguments (see the corresponding multiple row sums in Table 1). This can be done from a neutral, student-, or teacher-related perspective. When asked about school performance assessment, students focus on different perspectives, depending on whether they perceive themselves (again) as pupils in the past or as future teachers.

5.1.2.1 Neutral perspective

The neutral perspective means that there is no focus on acting persons; rather, situational contexts or processes are addressed. On the one hand, this suggests the distancing of personal attitudes, but on the other hand, it also allows for ambiguity with regard to role acceptance (the switch between teachers' and students' perspectives):

> It would be worth a try to investigate how the quantity and quality of the contributions would change if students were assessed according to their classroom participation. (ID 016)

5.1.2.2 Student's perspective

Research has shown how students hold on to their role as school students, while at university they are expected to take a reflective distance and practice their role as teachers (Wenzl, 2022). In the following text passage, a student's perspective is illustrated:

> This was the case in my former class, too. Back then, we had three very good students, who mostly refrained from participating completely, because they wanted to understand and internalize the content already in the lesson. (ID 017)

5.1.2.3 Teacher's perspective

The teacher's perspective and the influence of experience and judgment become apparent in the next text passage:

> To understand personal subject knowledge and the related competences as a dynamic process is even more important than particular exam results to my mind. I support the use of different exam methods, provided that an exam result depicts at least a part of the present learning progress. (ID 018)

5.1.3 Results on modes of argumentation

The modes of argumentation concern the main focus of reasons put forward in decision-making processes. Such a focus can be causal-analytic, normative, descriptive, or effect-oriented.

5.1.3.1 Causal-analytic mode of argumentation

A causal-analytic mode of argumentation was found only 11 times within 107 analyzed text sequences, making it the least frequent mode (see the corresponding column sum in Table 1). This mode of argumentation is related to the causes or previous conditions of factors related to performance assessment. Furthermore, it is often related to certain keywords like 'since' or 'because' and to different kinds of knowledge. Most commonly used is knowledge gained from personal experience (experienced or observed) or knowledge gained from experience in the practical context of the teaching profession (common, professional knowledge). Scientific knowledge, meanwhile, is used least often.

The following text sequence serves as an example for the above-mentioned mode of argumentation:

> I think the statement is true up to a certain extent, since it is important to spend time with a topic on one's own to learn something; however, interaction and discussion are [also] a great opportunity to learn. Everybody has different approaches and experiences which can be brought up for discussion in a dialogue and can help somebody else when studying the subject. In a lesson there are projects and group work too, where one can see how well "chatting" helps with studying. (ID 019)

5.1.3.2 Normative mode of argumentation

The normative mode of argumentation is related to more or less general rules or systems of rules and occurred comparatively frequently (31 times). This mode is used in argumentations that depict, in the first dimension, tensions between the school's claim of objectivity and the teacher's requirement to make decisions as well as those between a school's institutional system of performance assessment and students' learning processes. Those tensions are somehow presented as dilemmatic decision-making-situations for teachers. The second dimension can be summarized as the pedagogical ethos of quality ('A good teacher should …'), while the third dimension is related to the expectations and norms of contemporary society:

> We live in an era in which pedagogical work should not be measured by standardized maxims. Instead, we should take the chance to help students evolve, starting from their individual basis. I can only approve of individual oral exams when the teacher's purpose is to pay undivided attention to single students. (ID 020)

5.1.3.3 Descriptive mode of argumentation

The descriptive mode of argumentation appeared 44 times. It considers descriptions of real situations in schools and classrooms that are relevant for performance assessment. Educational phenomena are depicted as realities, partly in connection to their consistency. Such real situations are, for example, related to the school system and abilities of students:

> Unfortunately, quiet students are often neglected, since they appear not to participate. But this is mostly not the case, because especially the quiet [students] listen cautiously and attentively and should not be underestimated. It needs to be added that they often know the solution, but do not dare to speak in front of the whole class. (ID 021)

5.1.3.4 Effect-oriented mode of argumentation

The effect-oriented mode appeared 21 times. Here, one can recognize an important dimension of belief regarding school assessment, namely the assessment's impact on students' learning and achievement. We found statements on the effect of certain forms of performance recordings and performance assessments on future student knowledge and student competencies as well as teacher competencies:

> If the assessment is in accord with the type of teaching, then participation in discussions is to be included in the grade. A teacher who notices which students make good or bad contributions can benefit from [it]. Nevertheless, a teacher who thinks the interest of a student is based on any oral contribution will not benefit from [it]. (ID 019)

5.2 Mixed quantitative and qualitative results on the combination of reasoning structures

This section focuses on the quantitative distributions of reasoning structures as well as on correlations resp. combinations of their elements. Table 1 presents the quantified results of our coding and depicts t frequencies of codes in triple combinations of subcategories. Univariate distributions of frequencies can be seen within the sums of the rows and columns. For example, we found 49 statements on individual reference norms and 29 statements on neutral reference perspectives.

We found the following ranking of reference norms according to their frequencies: Individual reference norms (49) seems to be more important than objective (28), ability (21), and social (9) reference norms. From a quantitative resp. statistical perspective, this distribution is significantly different from a theoretically assumed uniform distribution with equal frequencies in each reference norm (based on a one-sample Kolmogorov–Smirnov-Test: K-S (99%,

$107) = 0.46$, $p < 0.001$). The results also reveal that a teacher focus dominates the reference perspectives (70) in comparison to a neutral (29) and a student (8) focus which is again different from a uniform distribution (K-S (99%, 107) $= 0.65$, $p < 0.001$). It is also apparent that descriptive (44) and normative (31) modes of argumentation outweigh effect-oriented (21) causal-analytical (11) types significantly (K-S (99%, 107) $= 0.27$, $p < 0.001$). Our quantitative tests confirm that there is a high degree of variability with salient points of focus in the reasoning structures on performance assessment of pre-service teachers.

Four text units with different triple combinations are particularly illustrative because of their relative frequency. Moreover, those sample combinations are typical argumentation patterns. Due to the low frequency for the student perspective, only the neutral perspective and teacher perspective were chosen for the following examples:

The first text unit shows the combination of the normative mode of argumentation, the individual reference norm, and the neutral perspective. The main issue in this unit is the monotony in teaching methods in contrast to the teachers' variation of teaching methods in the classrooms. With regard to the individual reference norm, it is assumed that those teachers who vary their teaching methods give each student the opportunity to address with their individual strengths and weaknesses. The mode of argumentation is dominated by the normative mode, and it can be described as a normative statement containing recommendations expressed by the modal verb 'should.' Additionally, it is reinforced by the neutral perspective because it refers to what teachers should generally pursue in their everyday teaching:

> In general, I see the problem of these expressions in the rigidity of teaching methods. In every school there are these teachers who only do frontal teaching, group work etc., there is no variety of methods. Here it should go, however. The teachers should offer a variety to the pupils, so that everyone can live out its strengths and weaknesses. (ID 023)

The second text unit includes a combination of the effect-oriented mode of argumentation, the ability-oriented reference norm, and the neutral perspective. Here, it is not the students' needs that define the assessment norm, but rather the external societal expectations of the required qualifications in the teaching profession. The mode of argumentation indicated by 'several perspectives' refers to the necessity to reflect teaching experiences from different ankles. Thus, the argumentative intention in this text unit presumably considers the potential impact of different performance assessments on prospective changes in society. The common setting of ability-oriented norm and effect-oriented mode of reasoning is written from a neutral perspective, which, in any case, refers to both students and teachers:

> In order to be able to answer such a question to students, it is necessary to consider the question from several perspectives. First of all, it should be mentioned that within the framework of the current curriculum, students are supposed to acquire a wide variety of competencies in the subjects. The three most important competencies are factual, social and self-competence. These competencies also play a central role in the further life of a student and are encountered by adolescent students particularly in everyday working life. Social competence is of great importance here. (ID 013)

The third text unit shows the combination of the descriptive mode of argumentation, the objective reference norm, and the teacher's perspective. The statement is clearly marked by the writer's identification with the teacher 'I do not evaluate'. The objective reference norm is described by clear guidelines and procedures and dominates the mode of argumentation, and it is closely linked to a descriptive mode of argumentation: The outlined behavior of the teacher in the classroom is described as a predetermined procedure that does not allow any other choice:

> I do not evaluate the frequency of words, but their quality. A question about what happens from the transition from the 7th to the 8th line does not lead me to conclude interest on the part of the student, but the lack of comprehensibility. So, my response would be to explain the step again. Verbal responses that are only questions for re-explanation are not counted toward collaboration, but rather serve as direct feedback to me as the teacher, telling me I need to re-explain. Oral collaboration must be assessed qualitatively, I cannot conclude good collaboration from the number of times I speak. (ID 024)

The fourth text unit shows the combination of the causal-analytic mode of argumentation, individual reference norm, and the teacher's perspective. When the writer refers to 'my students', a clear identification with the teacher's perspective becomes obvious. The causal-analytic rationale is twofold: Not only does the testing situation demand a differentiated perspective on the students' preconditions; it also requires an equal consideration of different performance assessments as well:

> I would give my students the answer that I feel it is very important to include oral contributions in the grade as well. Not only for the one reason that written performance reviews can often create a stressful situation in students' minds, but also because I feel that both oral and written performances should be graded equally. (ID 022)

Such examples represent combinations of reference norm, reference perspective, and mode of argumentation from a qualitative perspective. From a quantitative perspective, based on the data from Table 1, and by using chi-squared tests calculated in SPSS version 27, we found there to be some preliminary indication that there is a significant but weak relationship between reference norm and modes of argumentation ($(9) = 24.69$; Cramer's-V $= 0.28$, $p < 0.05$; however: 7 cells have an expected frequency < 5, minimal expected frequency: 0.93). Thus, this result needs to be interpreted with caution (particularly given that statistical preconditions on expected frequency are not perfectly met). The findings indicate that the effect-oriented modes of argumentation correspond with ability-oriented reference norms, whereas individual and objective reference norms are more frequent when using normative or descriptive modes of argumentation. We did not find any other statistically significant relationship between reference norm and reference perspective ($(6) = 3.17$; Cramer's-V $= 0.12$, $p > 0.05$) as well as between reference perspective and modes of argumentation ($(6) = 8.38$; Cramer's-V $= 0.20$, $p > 0.05$). Overall, quantitative results could serve as the first preliminary evidence of the fact that the different elements of reasoning structures, which provided the focus of our study, are more

or less strongly related to each other. This variability indicates that cognitive structures and processes during performance assessment vary in flexibility or stability and therefore in related implicit cognitive structures.

5.3 Results of the secondary analysis on implicit structures

The secondary analysis aimed to go one step beyond analyzing modes of argumentation. Such implicit structures are transmitted through 'narratives' that orient social practice according to the understanding of social reality due to the documentary method (Bohnsack, 1999). In three steps of interpretation, we analyzed how the pre-service teachers' textual statements already propose possible solutions for the assessment dilemmas and related tensions. The documentary method differentiates between the first step of the analysis, focusing on WHAT the text tells us, and the second step of the analysis, going beyond content-related dimensions and asking HOW the narrative is constructed, thus revealing the implicit knowledge within a certain orientation framework.

By asking WHAT is the explicit meaning of the texts, we discovered that all 10 texts resolved the tension of performance recording (oral versus written) and performance assessment and the often-contrasting impact on both learning and objectivity. Here we distinguished between three positions: one group argues for standardized assessment; another group highlights the importance of individualization methods that emphasize students' strengths (or reduce their weaknesses); and a third group prefers mixed forms and considers this the best way to neutralize or mitigate the problem of objectivity, reliability, and validity.

Asking for the forms of narratives tells us more about the implicit constructions of the social reality of assessment dilemmas. While the texts explicitly argued for different forms of performance assessment to relieve tensions, they also demonstrated that the form of assessment is not as important as the students using their knowledge to develop their learning experience. If an exam is necessary, then it should be standardized to ensure objectivity. However, assessment is not the purpose of education itself, and it becomes irrelevant in comparison to students' highly individual learning experiences. Thus, it can be argued that the dimension of belief in relation to the assessment's impact on students' learning and achievement is most important and all tensions and dilemmas can be solved with regard to the only important criterion, namely the effects on learning, achievement and further success in life, as the following example demonstrates:

> Relying only on written grading does not adequately represent a student's full range of skills. In many situations, it is also important to demonstrate one's abilities verbally. However, objectivity must not suffer in any examination mode. More important than specific exam results, however, I evaluate the understanding of personal expertise and related competencies as a dynamic process. (ID 018)

Aiming to establish a pedagogical equilibrium (Loughran, 2019), pre-service teachers attempt to resolve the dilemma situations by using similar but different perspectives. The narratives do not treat dilemmas as a conflict; rather, they try to establish a means by which to avoid the conflict either by devaluing the importance of assessment (e.g., learning

is more important than testing; working life is the real test) and neutralizing the conflict or by eliminating the tensions themselves by allowing students to choose their examination method. Other pre-service teachers, meanwhile, make efforts to close every possible gap between learning and assessment, viewing assessment as an instrument to help students prove themselves in everyday life and work.

The secondary analysis enables us to detect and extract the latent structures behind the participants' reasoning structures that are involved in the practical construction of social reality in the performance assessment dilemma. From a quantitative perspective and in order to identify implicit or latent factors within the reasoning structures, we have transformed all sub-categories into 11 variables with dummy codings and computed an exploratory factor analysis in order to explore latent factors (together with varimax-rotation and factor loadings >0.50 as criteria for building factors). We found three latent factors (with a cumulative explained variance of 83.68%) consisting of the following sub-categories: factor 1 (represents the reference perspective) is built of a neutral and a teacher-oriented reference perspective (factor loadings = 0.96 and -0.96), while factor 2 (reference norm) consists of an individual and objective reference norm (factor loadings = 0.91 and -0.84). Factor 3 is related to the ability-oriented reference norm and effect-oriented mode of argumentation (factor loadings = 0.88 and 0.73). It represents a combination of one aspect of a reference norm and of a mode of argumentation. We interpret this factor as an evidence-based mode of argumentation because effects as well as abilities or competences are strongly related to evidence-based forms of decision-making in our schools. Based on our criteria, we had to exclude the other sub-categories from our analysis. The eliminated sub-categories might be related to different or multiple factors simultaneously.

It might also be possible that these subcategories are of less importance resp. significance in daily decision-making on performance assessment than the other subcategories. Overall, our quantitative analyses did not confirm the findings on qualitative data, but they did not provide some indication that implicit latent structures may underlie pre-service teachers' reasoning structures on performance assessment.

6 Discussion

There is a growing interest in pre-service teachers' judgments regarding performance assessments, very often embedded within assessment literacy and novice teacher learning (Rogers et al., 2022). The present study explored the reasoning structures that pre-service teachers use in assessment dilemma situations, as well as their interconnections with each other, whilst also exploring which implicit structures served as the backbone when being confronted with the tensions of performance assessment dilemmas.

The quantitative analysis of reference norms in this study pointed to a clear dominance of the individual norm and emphasized pre-service teachers' beliefs about the importance of the individual reference norm for teachers' actions. However, teachers' beliefs regarding the individual reference norm do not necessarily correspond with their actual proceedings in practice, as they tend to primarily use the social reference norm (Dickhäuser et al., 2017; Marksteiner et al., 2021). According to school law, in turn, assessment should only be based on the objective reference norm, which follows the educational goals formulated in the curriculum (Neuweg, 2019). This could give

rise to curricular impulses in teacher education such as offering courses that encourage pre-service teachers to reflect on the contradictions between their self-assessment, teaching practice and school law. Furthermore, normative and descriptive modes of argumentation dominate and appear in combination with each other. Analytic and effect-oriented modes also occur together, but less frequently. This result suggests that analytical argumentation may not have the status among pre-service teachers that it ought to be given the nature of their profession. Consequently, recommendations for action can also be made for curricular practice, namely by paying more attention to fostering analytical argumentation skills in teacher education.

Given that the participating pre-service teachers have already completed their first internships, it is not surprising that the teacher's perspective dominates in this study. However, this perspective is often combined with a distanced view on teachers' actions. Specific combinations of chosen reference norms and modes of argumentation can be found more often than others. It is obvious, for instance, that certain epistemic views of performance assessment are more often connected to certain argumentation structures than others. The individual reference norm, which is argued from the teacher's perspective and from the neutral mode of argumentation, is far more commonly combined with normative and descriptive argumentations than with causal-analytic and effect-oriented argumentations. This may go hand in hand with the strong normative power of the argument in favor of the students' perspectives and assessment of fairness, but it would have to be explored in more detail in further studies. Furthermore, tensions between student-related individualization and factual or subject-related standardization in the presentation of achievement are mentioned. If the tension is addressed explicitly in a way that reference is made to several perspectives, including the quality criteria of performance appraisals, analytical arguments will increase while normative arguments will decrease.

The reconstructive analysis showed more clearly that the typical dilemmas of the performance assessment situation in schools, both theoretically and systematically, can be argumentatively balanced and interpreted in different ways.

According to our small sample, the participants tried to avoid presenting the dilemma as either a conflict and or as an unsolvable problem in their arguments, but rather relativized its importance by various methods or formulate solutions to restore a so-called pedagogical balance and thus the ability to continue acting in the classroom.

In future research, in-depth detailed analyses using the 'thinking aloud protocol' method could be suitable for identifying implicit beliefs and unexplained knowledge more precisely by asking the students about the subject of performance assessment. In educational research, the epistemological beliefs of actors/players/participants are typically recorded in a standardized way by using self-assessments and are relatively rarely collected via argumentation structures (Wu and Tsai, 2011), as these have to be made accessible by means of linguistic elaborates. This study taps into the growing interest in research into the perspective change from 'teacher thinking' to 'teacher writing' (Bullough, 2015). By using a topic-related text impulse that addresses the critical event of 'performance assessment' as an essential dimension of teacher action, this study attempts to use the method of reflective writing specifically for the exploration of argumentation structures, which, in turn, uncovers beliefs.

The results of this study also need to be interpreted in light the research's limitations. First, the chosen text impulse focuses on selected

problems of school performance assessment, which entail a certain thematic restriction. In the selected text impulse, this is certainly the dichotomous (conceptual) pair of oral vs. written performance assessment that appears in the text. It is clear that other text impulses with different instructions might have produced different results. Second, another limitation results from the varying text production competence of pre-service teachers in our sample, whose variation as an influencing factor was not considered. Third, we chose a highly cognitivistic perspective. We know from the literature that teaching and performance assessment often occur under pressure, where only superficial or faulty decision-making take place and no deep and flawless cognitive structures are relevant, as we suspected in our work. Future research must therefore focus in more detail on the errors and non-cognitive factors of the assessment processes of pre-service teachers (Astleitner, 2020). Fourth, the participants in this study were all enrolled in their fourth term with preliminary teacher experiences in performance situations. Consequently, future research could investigate the judgments of assessment dilemma situations encountered by pre-service teachers who have been exposed to more performance assessment situations during their internships. Fifth, we have conducted a mixed-methods analysis in which we combined qualitative and quantitative approaches. Our quantitative analysis confirms the qualitative results at least partially in the sense that (a) we found strongly varying reasoning structures, (b) we identified only some correlational patterns between reference norms, reference perspectives, and modes of argumentations, and (c) we were able to identify some latent factors behind the measured reasoning structures.

Despite these limitations, this study provides important insights into how pre-service teachers think and the dilemmas they must face in everyday classroom situations. We also provide some evidence that teacher training is less ineffective than some critics believe (e.g., Whitford et al., 2018). Indeed, if knowledge and reasoning structures vary, then they can be changed. If they can be changed, then they can also have a positive impact on everyday school life.

Author's note

Data availability statement

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

Author contributions

Acknowledgments

Conflict of interest

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2024.1170118/full#supplementary-material

References

Aguirre, J., and Speer, N. M. (2000). Examining the relationship between beliefs and goals in teacher practice. *J. Math. Behav.* 18, 327–356. doi: 10.1016/S0732-3123(99)00034-6

Allal, L. (2013). Teachers' professional judgement in assessment: a cognitive act and a socially situated practice. *Assess. Educ.* 20, 20–34. doi: 10.1080/0969594X.2012.736364

American Educational Research Association American Psychological Association and National Council on Measurement in Education (2014). Standards for educational and psychological testing. Available at: https://www.apa.org/science/programs/testing/standards (Accessed November 25, 2023).

Amrhein-Kreml, R., Bartosch, I., Breyer, G., Dobler, K., Koenne, C., Mayr, J., et al. (2008). *Prüfungskultur. Leistung und Bewertung (in) der Schule. [Exam culture. Performance and assessment at/of schools.] Institut für Unterrichts-und Schulentwicklung.* Klagenfurt: IUS.

Astleitner, H. (2020). "A theoretical perspective on ineffective interventions: malfunctions in teaching" in *Intervention research in educational practice.* ed. H. Astleitner (Münster: Waxmann), 39–62.

Baniasadi, A., Salehi, K., Khodaie, E., and Noaparast, K. B. (2023). Fairness in classroom assessment: a systematic review. *Asia Pac. Educ. Res.* 32, 91–109. doi: 10.1007/s40299-021-00636-z

Barnes, N., Fives, H., and Dacey, C. M. (2015). "Teachers' beliefs about assessment" in *International handbook of research on teachers' beliefs.* eds. H. Fives and M. G. Gill (New York: Routledge), 284–300.

Bendixen, L., Dunkle, M. E., and Schraw, G. (1994). Epistemological beliefs and reflective judgement. *Psychol. Rep.* 75, 1595–1600. doi: 10.2466/pr0.1994.75.3f.1595

Beziat, T. L., and Coleman, B. K. (2015). Classroom assessment literacy: evaluating pre-service teachers. *Researcher* 27, 25–30.

Bohnsack, R. (1999). *Rekonstruktive Sozialforschung: Einführung in Methodologie und Praxis Qualitativer Forschung [Reconstructive social research. Introduction to methodology and practice of qualitative research].* 4th Edn Wiesbaden: Springer.

Bohnsack, R. (2017). *Praxeologische Wissenssoziologie [Praxeological sociology of knowledge]* Opladen, Toronto: Budrich UTB.

Bortz, J., and Lienert, G. A. (2003). *Kurzgefasste Statistik für die klinische Forschung. Leitfaden für die verteilungsfreie Analyse kleiner Stichproben [Summary statistics for clinical research. Guide to nonparametric analysis of small samples]* Heidelberg: Springer.

Breidenstein, G. (2018). "Das theorem der Selektionsfunktion der Schule und die praxis der Leistungsbewertung the theorem of the selectional function of school and the practice of performance assessment" in *Leistung als Paradigma. Zur Entstehung und transformation eines pädagogischen KonzeptsPerformance as paradigm. Development and transformation of an educational concept*. eds. S. Reh and N. Ricken (Wiesbaden: Springer), 307–328.

Bullough, R. V. (2015). "Methods for studying beliefs" in *International handbook of research on teachers' beliefs*. eds. H. Fives and M. G. Gill (New York: Routledge), 150–170. Available at: https://books.google.at/books/about/Educational_Measurement.html?id=P4LuAAAAMAAJ&redir_esc=y

Camilli, G. (2006). "Test fairness" in *Educational measurement (4th)*. ed. R. L. Brennan (Praeger), 221–256.

Campbell, C. (2013). "Research on teacher competency in classroom assessment" in *SAGE handbook of research on classroom assessment*. ed. J. H. McMillan (Los Angeles, London: Sage), 71–84.

Cohen-Sayag, E., and Fischl, D. (2012). Reflective writing in pre-service teachers' teaching: what does it promote? *Aust. J. Teach. Educ.* 37, 20–36. doi: 10.14221/ajte.2012v37n10.1

Cole, N. S., and Zieky, M. J. (2001). The new faces of fairness. *J. Educ. Meas.* 38, 369–382. doi: 10.1111/j.1745-3984.2001.tb01132.x

DeLuca, C., Coombs, A., LaPointe-McEwan, D., and Chalas, A. (2018). Changing approaches to classroom assessment: An empirical study across teacher career stages. *Teach Teach Educ* 71, 134–144. doi: 10.1016/j.tate.2017.12.010

Dickhäuser, O., Janke, S., Praetorius, A.-K., and Dresel, M. (2017). The effects of teachers' reference norm orientations on students' implicit theories and academic self-concepts. *Zeitschrift für Pädagogicshe Psychologie* 31, 205–219. doi: 10.1024/1010-0652/a000208

Dollase, R. (2004). Was macht erfolgreichen Unterricht aus? [What makes successful teaching?]. Available at: http://www.finland.de/dfgnrw/doku/RainerDollase.pdf

Edmonds, W. A., and Kennedy, T. D. (2013). *An applied reference guide to research designs. Quantitative, qualitative, and mixed methods* Los Angeles, London: Sage.

Eggert, S., Ostermeyer, F., Hasselhorn, M., and Bögenholz, S. (2012). Socioscientific decision-making in the science classroom: the effect of embedded metacognitive instructions on students' learning outcomes. *Educ. Res. Int.* 2013, 1–12. doi: 10.1155/2013/309894

Fang, S. C., Hsu, Y. S., and Lin, S. S. (2019). Conceptualizing socioscientific decision-making from a review of research in science education. *Int. J. Sci. Math. Educ.* 17, 427–448. doi: 10.1007/s10763-018-9890-2

Fives, H., Barnes, N., Buehl, M. N., Mascadri, J., and Ziegler, N. (2017). Teachers' epistemic cognition in classroom assessment. *Educ. Psychol.* 52, 270–283. doi: 10.1080/00461520.2017.1323218

Fives, H., and Gill, G. M. (2014). *International handbook of research on teachers' beliefs* New York: Routledge.

Flick, U. (2013). *Triangulation. Eine Einführung [Triangulation. An Introduction]* Wiesbaden: VS Verlag für Sozialwissenschaften.

Franco, G. M., Muis, K. R., Kendeou, P., Ranellucci, J., Sampasivam, L., and Wang, X. (2012). Examining the influences of epistemic beliefs and knowledge representations on cognitive processing and conceptual change when learning physics. *Learn. Instr.* 22, 62–77. doi: 10.1016/j.learninstruc.2011.06.003

Frey, B. B., and Schmitt, V. L. (2007). Coming to terms with classroom assessment. *J. Adv. Acad.* 18, 402–423. doi: 10.4219/jaa-2007-495

Gholami, K., and Husu, J. (2010). How do teachers reason about their practice? Representing the epistemic nature of teachers' practical knowledge. *Teach. Teach. Educ.* 26, 1520–1529. doi: 10.1016/j.tate.2010.06.001

Gläser, J., and Laudel, G. (2010). *Experteninterviews und qualitative Inhaltsanalyse als Instrument rekonstruktiver Untersuchungen [Expert interviews and qualitative content analysis as instrument of reconstructive investigation]*. *4th* Edn Wiesbaden: VS Verlag.

Gogo, S., and Musondo, I. (2022). The use of the exploratory sequential approach in mixed-method research: a case of contextual top leadership interventions in Construction H&S. *Int. J. Environ. Res. Public Health* 19:7276. doi: 10.3390/ijerph19127276

Guerriero, S. (2017). *Pedagogical knowledge and the changing nature of the teaching profession* Paris: OECD Publishing.

Halkes, R. (1981). The assessment cycle. Towards a continual attentiveness to assessment and judgement in (teacher) education and evaluation. *Revue ATEE J.* 4, 77–89. doi: 10.1080/0379606810040104

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

Hofer, B. K., and Pintrich, P. R. (1997). The development of epistemological theories: beliefs about knowledge and knowing and their relation to learning. *Rev. Educ. Res.* 67, 88–140. doi: 10.3102/00346543067001088

Hofmann, F., Schmidinger, E., and Stern, T. (2016). "Leistungsbeurteilung unter Berücksichtigung ihrer formativen Funktion" in *Nationaler Bildungsbericht 2015*. eds. M. Bruneforth, F. Eder, K. Krainer, C. Schreiner, A. Seel and C. Spiel (Graz: Leykam), 59–94.

Ioannou-Georgiou, S., and Pavlou, P. (2003). *Assessing young learners* Oxford: Oxford University Press.

Jones, S., and Lawson, A. (2018). "Moving from projects to programmatic aid" in *Making development work*. eds. N. Hanna and R. Picciotto (New York: Routledge), 187–204.

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

Konstantinidou, A., and Macagno, F. (2012). Understanding students' reasoning: argumentation schemes as an interpretation method in science education. *Sci. Educ.* 22, 1069–1087. doi: 10.1007/s11191-012-9564-3

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

Lombard, M., Snyder-Duch, J., and Bracken, C. C. (2002). Content analysis in mass communication: assessment and reporting of intercoder reliability. *Hum. Commun. Res.* 28, 587–604. doi: 10.1111/j.1468-2958.2002.tb00826.x

Loughran, J. (2019). Pedagogical reasoning: the foundation of the professional knowledge of teaching. *Teach. Teach.* 25, 523–535. doi: 10.1080/13540602.2019.1633294

Loughran, J., Keast, S., and Cooper, R. (2016). "Pedagogical reasoning in teacher education" in *International handbook of teacher education*. eds. J. Loughran and M. Hamilton (Singapore: Springer), 387–421.

Maggioni, L., and Parkinson, M. M. (2008). The role of teacher epistemic cognition, epistemic beliefs, and calibration in instruction. *Educ. Psychol. Rev.* 20, 445–461. doi: 10.1007/s10648-008-9081-8

Marksteiner, T., Nishen, A. K., and Dickhäuser, O. (2021). Students' perception of teachers' reference norm orientation and cheating in the classroom. *Front. Psychol.* 12:614199. doi: 10.3389/fpsyg.2021.614199

Mayring, P. (2015). *Qualitative Inhaltsanalyse: Grundlagen und Techniken [qualitative content analysis: foundations and techniques]*. Weinheim, Basel: Beltz.

Mayring, P., and Gläser-Zikuda, M. (2008). *Die praxis der Qualitativen Inhaltsanalyse [the practice of qualitative content analysis]*. Weinheim, Basel: Beltz.

McMillan, J. H. (2011). *Classroom assessment: Principles and practice for effective standards-based instruction* Boston: Pearson.

Medjedovic, I. (2014). *Qualitative Sekundäranalyse [qualitative secondary analysis]* Wiesbaden: Springer VS.

Menz, C., Spinath, B., and Seifried, E. (2021). Where do pre-service teachers' educational psychological misconceptions come from? *Zeitschrift für Pädagogische Psychologie* 35, 143–156. doi: 10.1024/1010-0652/a000299

Mukherjee, S. P., Sinha, B. K., and Chattopadhyay, A. K. (2018). *Statistical methods in social science research* Singapore: Springer.

Neuweg, G. (2014). "Das Wissen der Wissensvermittler: Problemstellungen, Befunde und Perspektiven der Forschung zum Lehrerwissen [the knowledge of knowledge transmitters: problems, findings, and perspectives of research about teacher knowledge]" in *Handbuch der Forschung zum Lehrerberuf*. ed. E. Terhart (Münster: Waxmann), 583–614.

Neuweg, G. (2019). *Kompetenzorientierte Leistungsbeurteilung. Pädagogische und rechtliche Hilfestellungen für die Schulpraxis [competence-oriented performance assessment. Pedagogical and legal assistance for practice at school]* Linz: Trauner.

Ogan-Bekiroglu, F., and Suzuk, E. (2014). Pre-service teachers' assessment literacy and its implementation into practice. *Curric. J.* 25, 344–371. doi: 10.1080/09585176.2014.899916

Oxford University Press (2023). Dilemma. Available at: https://www.oxfordlearnersdictionaries.com/definition/english/dilemma?q=dilemma (Accessed November 25, 2023).

Penso, S., and Shoham, E. (2003). Student teachers' reasoning while making pedagogical decisions. *Eur. J. Teach. Educ.* 26, 313–328. doi: 10.1080/0261976032000128166

Perry, W. G. (1981). "Cognitive and ethical growth: the making of meaning" in *The modern American college*. ed. A. Chickering (New York, London: Jossey-Bass), 76–116.

Pope, N., Green, S. K., Johnson, R. L., and Mitchell, M. (2009). Examining teacher ethical dilemmas in classroom assessment. *Teach. Teach. Educ.* 25, 778–782. doi: 10.1016/j.tate.2008.11.013

Rädiker, S., and Kuckartz, U. (2019). *Analyse qualitativer Daten mit MAXQDA. Text, Audio und Video [Analysis of qualitative data with MAXQDA. Text, audio and video]* Wiesbaden: Springer.

Reh, S., and Ricken, N. (2018). *Leistung als Paradigma [Performance as paradigm]* Wiesbaden: Springer.

Reischl, J., and Plotz, T. (2020). "Documentary method" in *Analyzing group interactions. A guidebook for qualitative, quantitative and mixed methods*. eds. M. Huber and D. Froehlich (London: Routledge), 49–59.

Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment: a qualitative study. *Teach. Teach. Educ.* 27, 472–482. doi: 10.1016/j.tate.2010.09.017

Reusser, K., and Pauli, C. (2014). "Berufsbezogene Überzeugungen von Lehrerinnen und Lehrern [Job-Related Beliefs of Teachers]" in *Handbuch der Forschung zum Lehrerberuf*. eds. E. Terhart, H. Bennewitz and M. Rothland (Münster: Waxmann), 642–661.

Rogers, A. P., Mitescu Reagan, E., and Ward, C. (2022). Preservice teacher performance assessment and novice teacher assessment literacy. *Teach. Educ.* 33, 175–193. doi: 10.1080/10476210.2020.1840544

Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: a critical review of research. *J. Res. Sci. Teach.* 41, 513–536. doi: 10.1002/tea.20009

Schäffer, B. (2020). "Typologien als Endprodukt von Prozessen Typenbildender interpretation [Typologies as final product of processes of type-creating interpretation]" in *Jahrbuch Dokumentarische Methode (Issue 2–3)*. eds. S. Amling, A. Geimer, S. Rundel and S. Thomsen (Berlin: Centrum für qualitative Evaluations-und Sozialforschung), 23–48.

Schmidinger, E., Hofmann, F., and Stern, T. (2015). "Leistungsbeurteilung unter Berücksichtigung ihrer formativen Funktion [Performance assessment under consideration of their formative function]" in *Nationaler Bildungsbericht Österreich [National Educational Report]*. eds. M. Bruneforth, F. Eder, K. Krainer, C. Schreiner, A. Seel and C. Spiel (Graz: Leykam), 59–94.

Schuck, S., Aubusson, P., Burden, K., and Brindley, S. (2018). *Uncertainty in teacher education futures: Scenarios, politics and STEM* Singapore: Springer.

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

Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., and Marino, J. P. (2019). Assessment of university students' critical thinking: next generation performance assessment. *Int. J. Test.* 19, 337–362. doi: 10.1080/15305058.2018.1543309

Shaw, V. F. (1996). The cognitive processes in informal reasoning. *Think. Reason.* 2, 51–80. doi: 10.1080/135467896394564

Shulman, L. (1987). Knowledge and teaching: foundations of the new reform. *Harv. Educ. Rev.* 57:1–23. 22. doi: 10.17763/haer.57.1.j463w79r56455411

Spooner-Lane, R., Buchanan, J., Jordan, K., Broadley, T., Wall, R., and Hardy, G. (2022). "Exploring the impacts of a teaching performance assessment on Australian initial teacher education programs" in *Reconstructing the work of teacher educators*. eds. T. Bourke, D. Henderson, R. Spooner-Lane and S. White (Singapore: Springer), 157–178.

Stobart, G. (2006). "The validity of formative assessment" in *Assessment and learning*. ed. J. Gardner (Los Angeles, London: Sage), 133–146.

Suurtamm, C., and Koch, M. J. (2014). Navigating dilemmas in transforming assessment practices: experiences of mathematics teachers in Ontario, Canada. *Educ. Assess. Eval. Account.* 26, 263–287. doi: 10.1007/s11092-014-9195-0

Terhart, E. (2014). *Handbuch der Forschung zum Lehrerberuf [handbook of research on the teaching profession]* Münster: Waxmann.

Toulmin, S. (2003). *The uses of argument* Cambridge University Press.

Wegner, E., Anders, N., and Nückles, M. (2014). Student teachers' perception of dilemmatic demands and the relation to epistemological beliefs. *Frontline Learn. Res.* 5, 46–63. doi: 10.14786/flr.v2i3.83

Wenzl, T. (2022). *Rekonstruktive Bildungsforschung jenseits vom Fall* Wiesbaden: Springer.

Whitford, D. K., Zhang, D., and Katsiyannis, A. (2018). Traditional vs. alternative teacher preparation programs: a meta-analysis. *J. Child Fam. Stud.* 27, 671–685. doi: 10.1007/s10826-017-0932-0

Wu, Y.-T., and Tsai, C.-C. (2011). High school students' informal reasoning on a socio-scientific issue: qualitative and quantitative analyses. *Int. J. Sci. Educ.* 29, 1163–1187. doi: 10.1080/09500690601083375

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. doi: 10.1080/09695940903565610

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

Zhang, Z., and Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Appl. Meas. Educ.* 16, 323–342. doi: 10.1207/S15324818AME16044