ZDM (2018) 50:555–570
https://doi.org/10.1007/s11858-018-0963-z

SURVEY PAPER

Assessment in mathematics education: responding to issues regarding methodology, policy, and equity

Guri A. Nortvedt[1] · **Nils Buchholtz**[1]

Accepted: 19 June 2018 / Published online: 23 June 2018

Abstract
In educational contexts, assessments may be designed to target students, preservice teachers, or teachers, either as individuals or as representatives of a group, and for a multitude of purposes. One key aim of assessment in mathematics education is to provide evidence that can be used to make decisions about or improve mathematics education, which then raises questions about which aspects of mathematical competence should be assessed—as well as how and for what purpose. This review paper addresses three related themes: (1) issues related to the assessment process and to the development of assessments that can validly assess mathematical competence in all its complexity; (2) issues related to educational policy and policy-making based on assessment data, in particular the reciprocal relationship between assessment and policy; and (3) issues related to equity, such as gender issues or the achievement gap between majority and minority students. Awareness of the relation between assessment, teaching, and learning is shown throughout the paper. Strong relationships between the three focus areas are found, that impact assessment validity and call for further development of assessment practices in mathematics education.

Keywords Assessment · Validity · Methodology · Policy · Equity

1 Introduction

Why have a *ZDM Mathematics Education* issue on assessment in mathematics education? The last three decades have seen an increasing focus on assessment (see, for instance, Wiliam 2007; Suurtamm et al. 2016). From a strong focus on examinations and knowledge-based assessments, assessments for learning, national tests, and international comparative studies have gradually entered the scene. Today a wide range of different assessment formats and purposes exist. Throughout the world, various forms of assessment are employed to elicit information that can be used to inform decisions about mathematics education at the individual, institutional, and national levels. Many scholars have claimed that assessment should be used primarily to improve learning (Black and Wiliam 2012; Hattie and Timperley 2007; Niss 1993; van den Heuvel-Panhuizen and Becker 2003). Mathematics tests may also be used for admission to higher education or to monitor or evaluate the effectiveness of schools or educational systems. In addition, assessment outcomes are widely used to inform policy-making and decisions about educational reform (Elstad et al. 2009; Newton 2007; Nortvedt 2018).

Students, teachers, policy-makers, and even researchers may have naïve and strong beliefs about the objectivity and validity of assessments, including the belief that a single test or observation can tell the truth about the achievements of students, teachers, or educational systems (Stobart 2008). The applied assessment and the purposes for which the data are to be used may not align well (Newton 2007), which might be the case even when assessment is used for admission to higher education or for policy-making.

Mathematics education as a research field is young compared with mathematics itself (Kilpatrick 2014) or education in general (Wiliam 2003). Thus discussions within the field of mathematics education are often influenced by discussions in the neighbouring disciplines. When we discuss assessment in mathematics education, the discussion is often based on insights from educational research in combination with our knowledge about mathematics education and beliefs about mathematics. That is, current discussions of assessment in mathematics education reflect ongoing

✉ Guri A. Nortvedt
g.a.nortvedt@ils.uio.no

[1] Department of Teacher Education and School Research, University of Oslo, PO Box 1099, Blindern, N0317 Oslo, Norway

discussions about the purpose of mathematics education in general. Past and current debates have demonstrated that academics within our field might not share a unified understanding of the purpose of mathematics education (Niss 2007); we might not agree on which aspects of mathematics are worth teaching, or how learners learn mathematics. Our views about mathematics also colour what we believe should be assessed and how such assessment should be done, in addition to which issues we tend to identify when discussing current and future aspects of assessment in mathematics education. The mathematics education research community has engaged in numerous debates about assessment instruments, procedures, and outcomes over the past few decades. For instance, it is much easier to assess students' calculation skills than it is to assess their problem-solving skills, and many teacher-made tests primarily comprise algorithmic tasks (Palm et al. 2011; Schoenfeld 2007). What should a 'good' test look like, and what should it assess? Many of the unresolved issues that have emerged over the past decade (e.g. Black and Wiliam 2005; Kaiser et al. 2017; Niss 2007; Suurtaam et al. 2016) still remain unanswered. These debates about the methodological and technical issues connected to assessment design and implementation are concerned with not only what we assess but also how we assess and what conclusions we can draw from our assessments. Thus, these debates also concern how assessment outcomes can and are used in decision-making.

We should note that these factors refer not only to the issues and challenges we have discussed briefly in this introduction but also to the possibilities that assessment provides in connection with mathematics education 'for all'. Thus, the assessment debate concerns equity issues in addition to methodology and policy. Strong relationships might exist between equity and how we assess. For instance, when low-socioeconomic status (SES) students are frequently reported to have lower achievement scores than those of high-SES students (OECD 2013a; Mullis et al. 2016), is it because the low-SES students have acquired less of the measured competence, or because of some artefact connected to the items, or the assessment itself? The goal of this paper is to discuss issues connected with each of these three areas separately; we do so on the basis of a selective review of existing research literature on assessment in mathematics education. The three focus areas may be described as follows.

1. *Methodological and technical issues in developing and conducting assessments, related both to what is assessed and how it is assessed*—that is, to the relationship between the purpose of the assessment and the assessment format. The discussion focusses on the four stages of the assessment process, which include frameworks, operationalisation, measurement, and validation. The discussions include both classroom assessment and external or large-scale assessment. This section is somewhat longer than the two following sections.
2. *Policy issues related to the interpretation, use, and misuse of assessment outcomes in policy development and the possible consequences for mathematics education*. This aspect includes a discussion of the reciprocal relationship between assessment and policy.
3. *Equity issues, including equity and social-justice issues, and what should be taken into account to develop fair assessments*. We use gender differences and issues related to assessing migrant students as illustrative examples in the discussion of possible consequences of the current assessment policies and practices.

The primary aim of this special issue is to move beyond traditional divides—such as large-scale versus classroom assessment, assessment at different levels, and targeting of different groups—and instead to discuss more fundamental issues related to assessment in mathematics education.

Each of the articles in this *ZDM* special issue addresses one or more of the three focus areas, as these areas have many strong connections, including between assessment formats and the opportunity for students from diverse backgrounds to demonstrate their competence, among others. The 13 articles represent novel perspectives on the three issues identified for this article, or they discuss how these issues have been treated within mathematics education research in general. We have included them all among the pool of papers, book chapters, and articles that we used for this review; many of them appear in more than one of the three sections.

2 Review procedures

Our review might be termed a state-of-the-art review, according to the categorisation of Grant and Booth (2009), who state that such reviews "tend to address more current matters in contrast to other combined retrospective and current approaches. [They] may offer new perspectives on issue or point out area(s) for further research" (p. 95, authors' additions). In line with Grant and Booth's description, we conducted an extensive search of current research literature on assessment in mathematics education (2000–2018). No formal quality assessments (i.e., formal categorisations) were performed on the research results. Rather, the purpose of the analysis has been to describe the current state of knowledge on assessment in mathematics education—particularly in relation to the three areas of concern we identified above—and to discuss priorities for future investigations and development.

2.1 Emergence of the three issues

Traditionally, classroom or teacher assessment has been discussed separately from large-scale or high-stakes assessment. For example, in the *Second Handbook of Research on Mathematics Teaching and Learning*, edited by Frank Lester (2007), the section on assessment comprises three chapters that focus on classroom assessment, high-stakes testing, and international large-scale testing. This traditional division between classroom and high-stakes assessment still very much exists, although efforts have been made to look past it. In the research literature, authors now look at overarching issues and *problematiques* related to all assessment formats and purposes. For instance, members of the Topic Study Group (TSG) 33 on assessment and testing in mathematics education from the 12th International Congress on Mathematical Education (ICME-12) in Seoul reflected on broad issues such as the development of assessment tasks in light of the complexity of mathematical thinking or the design of alternative assessment modes (e.g., online testing) in mathematics (Suurtamm and Neubrand 2015). In addition, the members of TSG 39 (Large-Scale Assessment and Testing in Mathematics Education) and TSG 40 (Classroom Assessment for Mathematics Learning) from the 2016 ICME-13 congress in Hamburg chose to work together to develop a Springer volume on assessment in mathematics education (Suurtamm et al. 2016). This development indicates that certain issues related to assessment in mathematics education are not connected to format or level but rather to more fundamental, underlying structures. Assessment format and purpose affect students who take the assessment differently and also affect the lessons that we can learn from implementing an assessment or from analyses of assessment outcomes. As a result, not only methodology but also equity and policy have emerged as key issues to be included in this review.

2.2 Procedures

After identifying the three areas (methodology, policy, and equity), we performed thematic searches to identify scientific articles, book chapters, or books that address fundamental aspects in relation to these three issues. Searches were made using Eric, Google Scholar, and ORIA[1], applying search words and phrases such as 'mathematics assessment' together with terms such as 'methodology', 'policy', and 'equity'. The searches returned a large number of research publications, and the two authors used their knowledge of the field to identify any sub-topics within each of the three areas (cf., Grant and Booth 2009). A snowballing technique was applied to follow discussions or themes that emerged from the articles that were initially identified. The purpose of our review was to provide a good overview of emerging issues rather than to discuss each issue in full detail. The review thus did not follow the procedures typically used in systematic reviews; searches were ended once sufficient research had been found to identify central ideas and issues, thus enabling each topic to be discussed thoroughly. According to Grant and Booth (2009), the state-of-the-art review is particularly well suited to identifying the research approaches and main characteristics of a topic. A careful reader will discover that some of the issues we discuss are not novel but have been discussed for a long time. In these cases, we use 'older' references (2000–2007) together with more recent ones.

In this review we have narrowed the scope of our analysis to the assessment of competence of students, preservice teachers, and teachers. In some parts of the review, we have not distinguished between the three groups, since similar issues emerged in the literature search (e.g., connected to assessment formats) for all groups. In the sections that focussed on policy, research on policy-making for primary and secondary education emerged more often than research on policy issues targeted towards mathematics teacher education; we found the same situation for research on equity. Although we did find some research on equity issues related to the assessment of preservice teachers, very little of this research focusses on issues connected to underserved groups.

3 Issues in methodology

The major methodological issues in assessment in mathematics can be related to the 'what' and 'how' questions of assessment. The 'what' question relates to the aspects of mathematical competencies that can be validly assessed, while the 'how' question regards the assessment format or method for measuring the competencies addressed in the 'what' question. Figure 1 shows the assessment process, which consists of several consecutive sub-processes that guide the development of assessments, both in large-scale studies and in classroom assessments (for an example, see OECD 2013c).

Methodological challenges may be found in each of the sub-processes: (1) in defining a framework or conceptualising the content to be assessed, (2) in operationalising the framework and developing assessment formats and items, (3) in implementing the measurement, and (4) in interpreting

[1] ORIA is a search engine supported by the Norwegian University Library that searches several external databases simultaneously (https://bibsys-almaprimo.hosted.exlibrisgroup.com/primo-explore/search?vid=UIO&sortby=rank&lang=no_NO). Searches are performed in a large number of library data bases such as the Springer data bases or Thomson Reuter data bases, ProQuest.

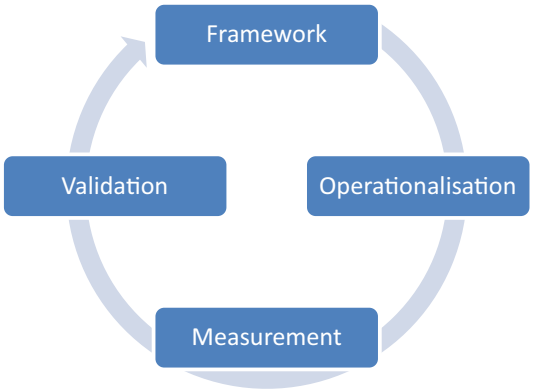

Fig. 1 The assessment process

the outcomes of assessment and validating the assessment instruments.

In educational assessments such as large-scale international and national assessments—exemplified by PISA, TIMSS, and national tests and examinations—an assessment framework is usually developed that describes the content or concepts to be assessed and how the framework should be operationalised. Standards regulating how the assessment should be implemented are often included in the framework or in accompanying guides or technical frameworks (e.g., OECD 2013c). In classroom assessments, a national curriculum may be viewed as an assessment framework, although a framework might also be provided by the local school or local school authorities that includes standards and goals for mathematics teachers. Teachers have to interpret these frameworks and standards and decide on what to teach and how to assess this content. They often use a range of teacher-selected or teacher-made assessment instruments that are closely linked to what the students have been learning (Suurtam et al. 2016; Wiliam 2007).

Teachers also need to interpret and validate assessment outcomes, both from the assessments they themselves make and from external assessments. If the tasks used for instruction and assessment are too similar, however, then there is a danger of overestimating what students have achieved, because the curriculum and instruction might be narrowed towards tested topics and even towards certain problematic styles or formats (Hamilton et al. 2007). Previous research has revealed that teacher-made tests often assess what might be termed lower-order skills (see, for instance, Palm et al. 2011). Even established large-scale assessments might test only certain aspects of mathematical competence; for instance the TIMSS study aims to assess what is shared by the curricula in the participating countries (Mullis et al. 2016). This lack of rigour might be a common issue shared at all levels of educational assessment. According to Niss (2007),

> very little progress has been made as regards the assessment of essential ingredients in mathematical competencies, such as asking questions, conjecturing, posing problems, constructing argument, including formal proofs, making use of and switching between representations, communication and such like. Not only is research lacking, assessment instruments are largely lacking as well. Much assessment and testing is still focused on students' solving of already formulated problems. This shows that as far as assessment is concerned there is indeed a long way to Utopia. (p. 1306).

In the years since Niss (2007) articulated this concern, some progress has been made in tasks for classroom assessment (as well as tasks in large-scale assessment) that better reflect the complexity of mathematical thinking and problem solving. Our field has even seen progress in how we assess various sub-areas of mathematical competence, including that of both students and preservice teachers (see, for instance, Suurtamm et al. 2016; Fujita et al. 2018; Ubuz and Aydin 2018). The PISA 2012 framework, for instance, attempted to describe and assess the different modelling and problem-solving processes, and identified these processes as the main reporting categories, rather than using the traditional division into different mathematical content strands as in previous cycles (Niss 2015; OECD 2013a, c).

But once we define what we will assess, how do we operationalise the framework and develop an assessment situation from it? Challenges connected to the process of designing and implementing an assessment will not differ depending on whom we assess, be it students in compulsory education or in teacher education; rather, it is the process of operationalising the framework content that is challenging for different stakeholders to agree on (Kaarstein 2014). A closer look at the research on mathematics teacher education might illuminate the relationship between framework conceptualisation and operationalisation. Neubrand (2018) provides an overview of different conceptualisations of mathematics teacher competence as well as approaches to assessing such competence. Starting from the assumption that teaching is a profession, he defines critical aspects of teaching competence and how major studies in the field have addressed these aspects over the past two decades. He concludes that a common denominator across projects is that they have disregarded teaching practice when operationalising the assessment framework. This disregard may be seen as a major weakness that affects the validity of these studies and calls for further methodological developments in mathematics (preservice) teacher assessments.

Further, we may identify the assumption of unidimensionality of content or competence defined in most assessment frameworks as a single trait as a key issue connected to the operationalisation of theoretical constructs defined

in corresponding frameworks. This assumption tends to yield large-scale assessments in particular (van den Heuvel-Panhuizen and Becker 2003). The conceptualisations and operationalisations that studies such as PISA make, subsume various mathematical activities that define mathematical literacy; they might include mathematising, arguing, proving, and problem solving, among others, under one broad psychological construct (e.g., OECD 2013c). Researchers often critique such assumptions when discussing large-scale studies in mathematics education (e.g., Wuttke 2007).

In what might be seen as a contrast to this approach, other studies might restrict what is measured in ways that could lead to construct underrepresentation. Several examples may be found regarding changes to how and what aspects of competence are assessed; these changes might seem like pragmatic choices at the time. For instance, when assessing teacher competence, we have observed a tendency to narrow the assessed abilities to discrete compartmentalised facets of teaching competence that are 'easier' to assess and can be explained with local theories from mathematics education, such as the teaching of algebra (e.g., Lynch and Star 2014), diagnostic competence (e.g., Hoth et al. 2016), or school-related mathematical knowledge (e.g., Buchholtz et al. 2013).

An important question to consider in assessment design is whether to attempt to assess an overall competence or whether to restrict what is to be assessed to some predefined aspects of teacher competence. For instance, Martinovic and Manizade (2018) describe in their paper the development of an instrument for assessing teachers' knowledge for teaching geometry. They focus on methodological issues connected to measuring mathematical knowledge for teaching; they also describe their approach to task design—targeting knowledge for teaching the area of a trapezoid and for accompanying assessment tools. Unlike assessing mathematics teacher competence on a more generic level, they discuss the benefits of developing assessment instruments within a well-defined and narrow topic in mathematics, and of combining different measures to ensure the validity of the assessed construct. This approach can provide insight into well-defined restricted areas of teacher competence. Still, questions remain of the generalisability of assessment results to other aspects of teacher competence, for example in terms of policy-making.

Another issue connected to operationalisation that applies to well-known teacher-education studies in mathematics is the lack of consideration of practical mathematical knowledge for teaching (Buchholtz et al. 2014). Most of the applied assessment frameworks to date neglect the teaching context that teachers experience in their classrooms. Thus the frameworks include only a few facets of professional abilities and lack generalisability to other teaching content or contexts, often overlooking what distinguishes elementary from secondary-level mathematics teaching (Rowland and Ruthven 2010; Speer et al. 2015). Corresponding issues of operationalisation can be identified regarding student assessment. Care must be taken when deciding on what content areas in mathematics (geometry, algebra, or arithmetic, for example) and what mathematical processes (such as proving, modelling, understanding, or interpreting) are to be assessed and linked to the theoretical concepts behind the traits that are to be assessed.

Moving on to the issue of measurement, assessment frameworks that integrate a wide range of mathematical processes or multiple tests, and where a variety of different assessment formats are involved, such as paper-based and computer-based testing, represent a methodological challenge, given that both assessment content and instruments in these cases are multifaceted and complex. Recent technological developments have facilitated a change of assessment mode, but these developments have not come without challenges. As of 2015, for example, the PISA study has switched from paper-based to computer-based assessment as its main assessment mode. Even if meta-analyses conclude that the mode of delivery does not greatly affect scores when assessing established constructs (Wang et al. 2007), other studies have revealed that factors such as on-screen reading, screen size, or resolution do affect cross-country comparisons (Jerrim 2016). One possible solution to this challenge of mode effects, which has been applied in the PISA study, is to statistically adjust for the results (OECD 2016). Still, a major question to be discussed is: Are we really measuring the same construct?

The issue of assessment mode relates to more than large-scale assessment. Several modes may be observed in classroom assessments that might involve a change from written tests to oral presentations or from paper-based to computer-based assessments. The shift from the extensive use of written tests to assessments for learning, for instance, might be seen as a shift from summative to formative assessment or from focussing on answers to focussing on mathematical processes (Wiliam 2007). Hoogland and Tout (2018) discuss how computer-based assessment might offer new opportunities to assess more of students' mathematical competence. For instance, recent developments in technology might support the assessment of higher-order thinking skills in mathematics while also offering opportunities to use authentic tasks. Fujita et al. (2018), for instance, discuss computer-based feedback and analyse the use of procedural feedback when conducting geometry proofs. In particular, they analyse how learners can overcome logical circularity with the help of computer-generated feedback and thus address more than issues about assessing procedural skills in mathematics, such as conducting a flow-chart proof in geometry. They discuss methodical challenges about developing computer-based assessments as well as the finding

that, for some students, supplementary teacher interventions are necessary, thus indicating that some further development is still needed.

A key difference between the approaches that Fujita et al. (2018) and Hoogland and Tout (2018) take compared to the OECD's approach (2016) is that the OECD, in PISA 2015, merely transferred their paper-based mathematics trend items[2] to a computerised format but failed to take advantage of the possibilities a digital platform might offer, as in the two other studies. While PISA for trend items primarily utilised possibilities for automated scoring, Fujita et al. and Hoogland and Tout discuss challenges related to designing richer and more authentic tasks. Hoogland and Tout state that more and more advanced and sophisticated tools exist that now enable efficient, automated testing and scoring of what might be termed lower-order content, such as calculation skills, procedural speed, and factual knowledge. Thus, technology can also limit what is assessed. The way forward might be to introduce frameworks that will allow categorisation of task content and, as such, could be used to scrutinise the operationalisation of the test content.

Another challenge in assessment design is to avoid compartmentalisation or the loss of detail due to overly inclusive assessment designs. This scenario might be remedied by combining or integrating diverse assessment formats in mathematics into the same assessment, or by applying several linked assessments to assess a larger variety of contexts or situations. When a single assessment is used in situations where far-reaching consequences might occur—such as admission to further studies or for certification—such situations involve a great risk of making the wrong decision, since affective and physical conditions can influence test takers' possibilities of demonstrating their competence (Pajares and Miller 1995; Ma 1999).

Learning or longitudinal development of mathematical thinking cannot be displayed by a single summative assessment. Correspondingly, in order to contribute to fair assessment, all assessments should take into account different and multiple sources of individual students' performance (Leder and Forgasz 2018), including classroom-based performance. High-quality formative and summative assessments, including multiple modes such as computer-based assessment, offer students a range of opportunities to demonstrate their mathematical knowledge (National Council of Teachers of Mathematics 2016). Buchholtz et al. (2018) discuss how an approach that integrates summative and formative assessment formats in mathematics teacher education might contribute to the preservice teachers' opportunities to learn in a German teacher education programme. All assessments were administered to a group of preservice teachers participating in a school practice course. In their paper, the authors integrate findings from an analysis of survey data with analyses of e-portfolio data to gather as much information as possible on preservice teachers' learning opportunities and the acquisition of situation-specific skills during the course. Leder and Forgasz (2018) also discuss how several assessment formats might be combined to create equal opportunities for male and female students to demonstrate their mathematical competence in a summative assessment. While the authors indicate that not simply using multiple tests but using multiple tests with different types of tasks and different formats might be more equitable, integrating or combining similar or different assessment formats might influence the validity of the conclusions that are drawn from the interpretations of assessment outcomes.

Any assessment needs to be validated so that all interpretations drawn from the assessment results will be justified and appropriate for the intended use and the assessed group of students or teachers (Newton and Shaw 2014). Pankow et al. (2018) present a specific validation approach for an assessment of teacher competencies. They validated a test to be used for assessing teachers' perception of students' errors: a capacity they judge to be crucial to mathematics teachers' competence. To check whether the test they developed was appropriate for the intended group, they compared the test results of different control groups, including students, preservice teachers, mathematics teachers, and mathematicians. They concluded that not only could mathematics teachers recognise students' errors faster than could the other groups, but their perception of students' errors was also more closely related to the domain of teachers' mathematical content knowledge than to the domain of teachers' mathematics pedagogical content knowledge.

Ubuz and Aydin (2018), who also addresses the validation of test instruments and the use of test results in educational research, applies the Standards for Educational and Psychological Testing (AERA, APA, and NCME 2014) to validate an assessment of students' knowledge of triangles. Unlike most tests, this test was developed to be multidimensional and to assess declarative, procedural, and conditional knowledge about triangles. Ubuz and Aydin applied factor analysis to identify differentiated structures between the different knowledge facets and thus broadened the validity of her instrument; she also stresses the need to take the external validity of assessments into account.

Validity not only affects different facets of test design but also affects the use of assessment data to inform educational decisions such as policy-making. In terms of both intended and unintended consequences, any inferences drawn from invalid conceptualised assessments might have far-reaching consequences, which lead to an important question: Are assessment data valid for what we want to use them for?

[2] This remark concerns only trend items. New science items used in PISA 2015 utilised the digital media to a larger extent.

4 Policy issues

According to Ernest (2014), policy in mathematics education is primarily related to the curriculum, that is, what is viewed as valid and important to teach in compulsory and further education. He distinguishes amongst policy aspects such as (1) policy debates about the aim and design of mathematics education, (2) those concerning mathematics teacher education, and (3) the assessment system. As such, policy influences mathematics education, both what happens in classrooms and what happens in teacher education. Although the term 'educational policy' usually refers to principles or guidelines that educational authorities have developed, policies might also be developed by a municipality, a school board, a school, or even a teacher association (e.g., Australian Association of Mathematics Teachers Inc. 2008). Such guidelines might identify how different assessments are to be used by different stakeholders in education (see for instance Elstad et al. 2009).

A key aim of educational assessment is to elicit evidence to be used to inform or monitor teaching. For instance, international comparative studies might provide the educational authorities of a country with insights about how their country is doing compared with others (Sälzer and Prenzel 2014). Assessment studies can provide useful information to improve mathematics education (Cai et al. 2016); thus assessment is a primary source for policy-making. Burkhardt and Schoenfeld (2018) argue that significant advances have been made within the field of mathematics education in conducting both formative and summative assessments but that these advances have not made a comparable impact on learning. In their review of previous research that has aimed to improve assessment practices and the use of assessment data to improve teaching and learning at the classroom level, the authors conclude that policy-makers typically underestimate the challenges involved in assessment design and development. This underestimation can lead to low-quality high-stakes tests and to a lack of uptake of assessments for learning in mathematics classrooms.

When educational policies are developed based on empirical data from assessment studies, this is usually referred to as evidence-based policy-making (Gaber et al. 2012). For instance, Hsieh et al. (2014) discuss how Taiwan used The Teacher Education and Development Study in Mathematics (TEDS-M) results to inform and adjust mathematics teacher education. Further, Lin et al. (2018) applied a four-phase framework that recognises social and cultural characteristics of both Taiwanese and Western educational systems to discuss how outcomes from international comparative studies have influenced educational policy in compulsory education and teacher education in Taiwan. Their analysis focusses on the chain of policy development, from assessment studies to an educational vision, followed by the implementation of a change to national assessments or curricula. Lin et al. to a large degree attributed the mechanisms of success and failure in evidence-based policy-making to traditional Confucian heritage, although they also found influences of Western thinking on curriculum development.

Applying assessment data to shape educational policy does come with certain risks connected to operationalising the theoretical constructs found in the assessment framework. This situation is illustrated in a study by Yang Hansen and Strietholt (2018), who reanalysed PISA data to investigate whether schooling perpetuates SES inequalities in mathematics performance. The rationale for their study is that previous research has shown that low-SES students show lower mathematics achievement than students from high-SES backgrounds. Yang Hansen and Strietholt, in noting that previous research has found that schooling does not contribute to learning to the same extent across ability levels, suggest that this situation is due to how opportunity for learning (OTL) is often measured. The complexity of the test design, as well as the analytical approaches used in international large-scale studies, together make it challenging to apply these studies for policy-making. A study by Shen and Tam (2008) illustrates this scenario very well. They examined the problem of culturally different reference standards by comparing subjective indicators with student performance in TIMSS data from 1995, 1999, and 2003. While measures of students' liking of mathematics and science, self-perceived competence, self-perception, and mathematics achievement were usually slightly positively correlated within countries, the correlation was negative when the authors conducted between-country analyses. This outcome was most likely due to the high academic standards attributed to high-achieving countries and the low academic standards attributed to low-performing countries.

Clearly, evidence-based policy is challenging, and evidence exists that stakeholders at all levels often reduce the complexity found in assessment data when looking for information to support their policies. For instance, rather than using the detailed information that assessments offer to identify possible insights or shortcomings in mathematics education, both the media and policy-makers tend to focus more on overall results and the ranking of countries; they conceive the rank as a quality indicator (Auld and Morris 2016; Hopfenbeck and Görgen 2017). Some efforts may be observed in which agencies and official bodies attempt to break down assessment outcomes to better explain and inform policy-makers about assessment outcomes; one example is the OECD report series *Education at a Glance*. Such efforts might not take into account the cultural context of national educational systems. Analyses by Baird et al.

(2016) and Nortvedt (2018) have indicated that the national context influences policy decisions to a large extent.

Many academics question the use of assessment data to inform educational practice (e.g., Biesta 2009) and voice a concern that large-scale international studies move the world towards a globalised and more uniform mathematics education that fails to take national traditions or needs into account (Stobart 2008). One might argue that such uses of international studies are based on surface-level analyses and that in-depth analyses of assessment data are still necessary. Consequently, the question of what might be done to improve stakeholders' abilities to interpret assessment data remains a key question. Burkhardt and Schoenfeld (2018) propose that the research community within mathematics education should engage with decision-makers. The authors indicate that the real challenge, however, is to find better ways to mitigate trust in simple statistics gleaned from what they have called high-stakes low-quality tests as well as increasing interest in high-quality assessments that can contribute to improved mathematics education.

Policy-makers and teachers alike often struggle to interpret the assessment data provided by high- and low-stakes tests and to use the test outcomes to inform their teaching (Groß Ophoff 2013). This scenario is often the case with external assessments such as international comparative studies and national tests, where what is tested often does not cover the full national curriculum, and only sample test items are released. In this situation, test outcomes might be used primarily for school self-presentation rather than to initiate change processes (Brown and Harris 2009). In addition, teachers might feel controlled rather than encouraged by the assessments (Stobart 2008). Hallinger and Heck (2010) promote collaborative school leadership, where school leaders and teachers share accountability for school practices (such as ownership and responsibility for assessment outcomes) and collaboratively interpret assessment data and plan interventions. Their research indicates that such practices can, over time, lead to higher student achievement in mathematics.

A key issue discussed in the research literature concerns the use of assessment data for policy development (Lin et al. this issue; Nortvedt 2018) and whether existing assessment practices provide the information that stakeholders need to make informed decisions (Gaber et al. 2012). This issue points to the potential consequences linked to the use of assessment data to inform decisions, since (1) more than one stakeholder is often involved and (2) the potential for misinterpretation always exists. Different stakeholders will apply assessment results for different purposes and therefore often need different kinds of evidence to support their decisions (Newton 2007). For example, some countries implement national tests to provide evidence both to teachers and decision-makers. That is, the same assessment should inform teaching and should provide information about individual students as well as information that can be used to evaluate the success of mathematics education at the regional or national level. Having multiple purposes such as these is highly problematic, since an assessment that validly and reliably provides information at the national level might not provide the same at the student level (Newton 2007; Fischer 2004).

To summarise, a two-directional, reciprocal relationship may be seen between policy and mathematics education, where assessment outcomes inform and influence policy-making (Baird et al. 2016) and educational policies influence assessment, teaching, or education programmes (Cai et al. 2015, 2016; Middleton et al. 2015). For instance, the phenomenon of 'teaching to the test' (Hamilton et al. 2007) is usually interpreted as a negative policy influence on mathematics education. In this case, low-performing students might be asked not to attend school on the day students are tested, or teachers might fabricate test results (Nichols and Berliner 2007). Equally, national examinations and high-stakes tests may influence the content that is offered within mathematics teaching (Hamilton et al. 2007). Educators have been discussing the 'what you test is what you get (WYTIWYG)' principle for a long time, but it is a principle that works both ways: educational authorities might initiate assessment reforms to influence mathematics teaching. While emphasising certain educational standards more than others can restrict the implementation of a curriculum, educational authorities might also use changes to national examinations to influence changes in mathematics education (Lin et al. 2018), including the uptake of digital tools such as dynamic geometry or CAS tools.

5 Equity issues

Equity in mathematics education concerns equal opportunities to learn important mathematical content for all students (see for instance Burkhardt and Schoenfeld 2018). Similarly, equity within mathematics assessment means that all students should have the same opportunities to demonstrate their mathematical competence. In an educational system focussed on 'education for all' (Niss 2007), equity is the gold standard compared with equality, where the same treatment is offered to all students but without the recognition that different students might need different kinds of support to achieve equity (Heritage and Wylie 2018). Achievement gaps between groups of students might indicate inequity, especially if the differences are systematic. Various gaps such as these exist today; for instance, both gender differences and an achievement gap between majority and minority students are frequently visible in mathematics assessments (e.g., OECD 2013a, 2015; Klenowski 2009). Indeed,

we could point to several cases of inequitable assessment practices, for example with regard to centralised versus decentralised national examinations (Wößmann 2005) or assessments of students with special needs (Scherer et al. 2016). In this review paper we use gender differences and the achievement gap between majority and minority students to illustrate how equity issues are linked to methodology and policy.

Gender differences in mathematics education have been discussed for a long time (Leder and Forgasz 2018). Achievement differences have traditionally been visible in large-scale studies and high-stakes achievement tests such as examinations and national tests (e.g., Liu and Wilson 2009; OECD 2013a). Male students usually outperform female students, although the reverse pattern is visible in some instances. Girls outperform boys in TIMSS or PISA in some countries for mathematics overall or for specific content areas (OECD 2013a; Mullis et al. 2016). Interestingly, the last PISA cycle showed that within OECD nations, the gender differences in mathematics are decreasing (OECD 2016), indicating that gender differences in achievement might not be particularly consistent across countries and time. An emerging alternative explanation concerns the mathematics teaching and curricula students are exposed to. Ayalon and Livneh (2013) found that in countries with a standardised educational system, boys and girls are exposed to the same content and teaching activities, which might lead to more similar achievement. The authors identified gender stratification as the most marginal factor involved in the creation of an achievement gap. We argue that these findings relate to the validity of the assessment, since the interpretations of achievement gaps might reflect national or institutional differences in schooling. When boys and girls are not exposed to the same curricula or teaching experiences, the same test might not be a valid assessment for both groups.

In addition, differences between male and female students have previously been visible in attitudes toward mathematics and beliefs about mathematics or oneself as a mathematics student (e.g., one's self-efficacy and motivation). Boys generally report more positive attitudes, while girls tend to report more anxiety (OECD 2013b). This difference might be an indication of varying attitudes or beliefs, or it might indicate gendered response patterns to survey questions. Indeed, some researchers have suggested that gender differences might be the outcome of different attitudes towards the test situation (e.g., test anxiety or performance avoidance) rather than real differences in mathematical knowledge or competence (Cotton et al. 2010; Hannon 2012; Hyde and Mertz 2009; Lindberg et al. 2010). Studies often explain gender differences in mathematics achievement by boys' more positive attitudes toward competition in general, but when controlling for such factors, these gender differences disappear (Hannon 2012; Cotton et al. 2010). The use of beliefs as a control variable might be challenging when investigating gender differences in achievement, given potentially different response patterns to questionnaires that investigate self-beliefs. If gendered ways of expressing oneself can be identified when beliefs are measured, such expression should be taken into consideration when reporting outcomes. Gendered ways of responding might also yield assessment formats. For example, Leder and Forgasz (2018) discuss how assessment format and purpose influence male and female students' assessment outcomes. Taking into account differences in interaction patterns, they conclude that within the Australian context, mathematics assessment still leads to inequity. The authors question whether national tests provide the same credible and important information about boys and girls; they ask critical questions about the terminology of testing and test validity from a gender perspective.

Equity is a key concern in multicultural classrooms. The heterogeneity of students—and consequently the diversity in mathematics classrooms—has increased considerably in recent years (OECD 2015), not least as a result of the increased number of refugees and the integration of students from crisis regions into the world's school systems (Paxton et al. 2011). Many refugees have very little formal education and severely interrupted or no substantive schooling, all of which limits their education (Miller and Mitchell 2006). Increased migration for work purposes is also visible in our globalising society. The PISA study has shown that 13% of the 15-year-old students in OECD countries came from immigrant backgrounds in 2015, compared with only 9% in 2006. In this context, classrooms are not only characterised by increased linguistic heterogeneity but also by increased heterogeneity in terms of prior mathematical knowledge (OECD 2016).

International comparative studies (e.g., OECD 2015; Museus et al. 2011) have frequently identified achievement gaps between majority and minority students. Similar patterns may be observed in national assessments, and overall differences between majority and migrant students might indicate inequity in relation to mathematics teaching and/or assessment. While first-generation immigrant students tend to score significantly below non-immigrant students in most countries, second-generation immigrant students tend to perform somewhere between the two groups (OECD 2015), although a fair degree of complexity makes the result patterns a challenge to disentangle. In PISA 2012, for instance, immigrant students scored at the level of non-immigrant students in some countries (e.g., New Zealand); in addition, second-generation students scored higher than non-immigrant students in a few countries, such as Australia (OECD 2015).

Previous research has identified language factors as a factor contributing to the assessment gap between majority and minority students (Klenowski 2009; Abedi and Lord

2001). Paper and pencil tests that assess students' mathematics competence might draw heavily on a student's ability to read and interpret texts. Previous research has demonstrated a high positive correlation between reading comprehension and mathematics achievement (e.g., Nortvedt 2011; Nortvedt et al. 2016). The issue of language as a 'gatekeeper' to accessing test content, as well as the consequences of not mastering the language of assessment to a sufficient extent, are also illustrated by the strong negative effect of late arrival (and consequently shorter exposure) to language instruction in the host country (OECD 2015). An alternative explanation for the achievement gap is the efficiency (or lack thereof) of the educational system of the host country. Migrant students from the same origin country tend to perform very differently in different educational systems; students from Arabic-speaking countries, for instance, performed far better in the Netherlands than they did in Finland on PISA 2012, although the average scores for the two countries were very similar (OECD 2015).

Finally, the gap between majority and minority students might also be partially due to how we assess students' competence or whose curricula we assess (Stobart 2008). Heritage and Wylie (2018), who present a case study on formative assessment in mathematics education, discuss a sample lesson taught by a teacher who has implemented assessment for learning in a multicultural classroom. Heritage and Wylie address the challenges and benefits connected to assessment for learning; in particular, they address identity and equity questions while highlighting the challenges that language requirements offer within the mathematics classroom. They conclude that effective assessment for learning practices can support both language and concept development among minority students, even when these students are instructed in a language they do not speak.

While we often assume that mathematics education is culture-neutral, research indicates that the way in which we express ourselves and view mathematics is in fact highly cultural (Klenowski 2009). A primary question about equity then relates to the opportunities that both majority and minority students have to use and display culture-specific knowledge in assessment situations, and whether this knowledge or competence is generally acknowledged as valid and important mathematical knowledge (Klenowski 2009). The lack of culturally responsive assessment could result in unequal educational opportunities for migrant students (Hopson and Hood 2005). The outcome could be that migrant students will be more likely to leave school early (Bradshaw et al. 2008) or that fewer migrant students will be admitted to higher education (Hopson and Hood 2005).

The concept of culturally fair or responsive assessment is challenging to define, as doing so necessitates a broad discussion of the term 'culture'. Montenegro and Jankowski (2017) describe culturally responsive assessment as being student-focussed by (1) being mindful of the student populations within the class/school/district/country, (2) using appropriate language, (3) acknowledging students' differences, and (4) applying tools that are appropriate for diverse students and that are applied with the intention of improving learning for all students who participate in the assessment. Care should be taken in all phases of assessment development and implementation to allow for a valid assessment of student competence, which could be taken into account by utilising tasks with fewer language barriers or introducing computer-based assessment. Computer-based assessment offers different means to contextualise and display mathematical tasks that are more connected to the realistic situation the task is supposed to present, such as datasets, video vignettes, graphical displays, and other means of presenting content and mathematical problems.

Task aspects such as video and graphical displays might also lessen the need for language fluency to comprehend the mathematical problem at hand (Sangwin 2013; Hoogland and Tout 2018). For example, in national low-stake tests, supplementary test items might be developed for students with special needs or language barriers in order to achieve higher measurement accuracy and to ensure that test items have appropriate difficulty levels and are generally understood by the students (Institut zur Qualitätsentwicklung im Bildungswesen IQB 2017). Students might also be empowered if they learn how to self-regulate, as this capacity is crucial to mathematics learning (Semana and Santos 2018). Students who participate in mathematics teaching with teachers who help them understand assessment criteria and participate in class discussions, will develop their capacity to self-regulate in mathematics learning situations, but to different degrees. Semana and Santos propose that some students might need more support than others to develop this capacity, and that differentiation is necessary to achieve equity with regard to students' opportunities to learn mathematics.

To judge if an assessment is fair, we need a theoretical framework comprising relevant factors. While some have advocated a socio-cultural perspective for framing assessments (e.g. Klenowski 2009), others have called for culturally responsive assessments (e.g. Montenegro and Jankowski 2017). Research literature focussed on how equity for migrant and majority students might be achieved in high-stakes testing remains scarce, however, as is culturally responsive mathematics assessment initiated by the teacher. In addition, previous research has revealed challenges for teachers in relation to using rich tasks in assessment situations (e.g., Siemon et al. 2004). Further, Wong and Glass (2005) identified challenges in designing professional development to sensitise teachers to culturally responsive assessment.

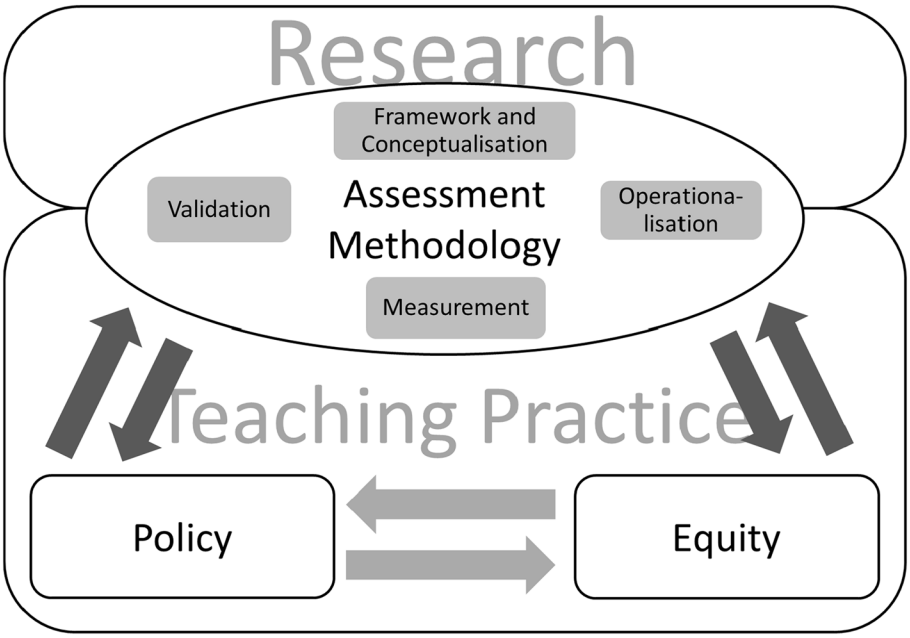

Fig. 2 Relations between assessment, policy, and equity

6 Issues in assessment in mathematics education

According to Niss (2007), 'mathematics for all' may be viewed as the goal of mathematics education, since it offers equal opportunities to develop mathematical competence to all students. Educational systems should educate citizens who can contribute to democracy and add to the technical and financial development of society; such citizens must also possess the mathematical competence necessary for their professional and everyday lives. To achieve these goals, we need to develop assessments that can assess all aspects of mathematical competence, not only certain aspects that are easier to assess (Hoogland and Tout 2018; Burkhardt and Schoenfeld 2018). Research in the field should pay more attention to the theoretical foundations of what is assessed and the development of the test instruments (Neubrand 2018; Martinovic and Manizade 2018; Yang; Hansen and Strietholt 2018; Ubuz and Aydin 2018; Pankow et al. 2018).

Much research and development remains to be done before equity issues in assessment in mathematics education can be properly dealt with. To better target assessments to individual levels of performance, we need more richness and variety in assessment formats (Leder and Forgasz 2018; Buchholtz et al. 2018). Not only culture and language need to be taken into consideration but also how students respond to feedback (Heritage and Wylie 2018; Fujita et al. 2018; Semana and Santos 2018). As the discussion in this review has shown, much still remains to be done 10 years after Niss's (2007) statement about current issues connected to assessing mathematical competence. The current assessment practices influence both methodological aspects and equity issues as well as the opportunity to use assessment for policy development. Ideally, the assessments used for policy-making should provide important information necessary to shape educational policies that can improve mathematics education (Lin et al. 2018). In addition, policy-makers must look past surface output (e.g., average scores) to identify the crucial messages (Auld and Morris 2016; Burkhardt and Schoenfeld 2018).

6.1 Relations between methodology, policy, and equity

In this section of the review, we do not treat issues related to methodology, policy, and equity as three separate issues because in reality, strong links are visible between the three areas of concern. Figure 2 displays the reciprocal relationships between methodology, equity, and policy. We should note that assessment is related both to teaching practice and to research. In the previous sections we identified relationships between the assessment methodology and policy and between assessment methodology and equity.

The reciprocal relationships displayed in Fig. 2 are visible also in the literature we have reviewed in this paper, where authors often discuss mathematics teaching and learning in relation to assessment and assessment outcomes, or in relation to policy implementation. Researchers might decide to conduct research on how different measures influence the inferences we may draw from assessments; they might also use output from international studies to inform educational authorities about changes to teaching, learning, and assessment in mathematics that could be included in new policies (Lin et al. 2018). In fact, a reciprocal relationship also exists

between research and practice, which reveals the complex relationships between the areas we have discussed in this review (see Fig. 2). What we assess, and how we assess, influence practice and policy. For instance, when large-scale studies reveal an achievement gap between migrant and majority students, that gap might influence both educational policy and equity, thus showing the connection between these two fields and how we develop or deliver assessments.

6.2 Assessment validity

The mutual influences of methodological, policy, and equity issues on one another are not least characterised by questions of assessment validity. Validity in general is the most fundamental, but also the most complex, quality criterion of any assessment. Despite the many debates about conceptions of different types of validity in the history of educational and psychological measurement (Newton and Shaw 2014), the current standards for educational and psychological testing portray validity as a 'unitary concept' (AERA, APA, and NCME 2014). Following this understanding, those in the field distinguish various sources of evidence that they might use to support the interpretation of assessment outcomes so that these interpretations are not only convincing and plausible but also empirically and methodically justified and acceptable to society. Sources for validity evidence (such as content or construct representativeness) affect methodological issues in several ways (Messick 1995). For example, the content of an assessment influences how valid the conceptualisation of important constructs in the assessment framework is, while the operationalisation of an assessment involves questions about meaningful and construct-valid tasks and test formats.

The validity of the measurement itself is also influenced by the technical quality of the assessment instrument as well as how the assessment is delivered to students, preservice teachers and teachers. Other sources of validity (such as the consequences of participating in an assessment) are more relevant to policy and equity debates because these validity sources are concerned with the interpretation of assessment results or plausible consequences of this interpretation to a greater extent. Validity thus relates to decisions made by various stakeholders at different levels in the educational system, from the teacher making informed decisions in the classroom, to the teacher educator or teacher education institution making decisions about teacher education programmes or professional development programmes, to policy-makers shaping a novel educational policy. The data these stakeholders apply come from different sources, from classroom discussions, observations, and teacher-made tests, to examinations and large-scale national assessments or international comparative studies. Validity arguments must take the quality of the data and any possible intentional or unintentional consequences of the argument into account.

Both intentional and unintentional consequences have previously been visible in educational systems that have strong accountability practices and external controls. In these cases, studies have found that teachers often teach to the test, low-performing children might be asked not to participate in assessments such as national tests, and teachers are often evaluated based on students' test scores (Baker et al. 2010; Seeley 2006). Although the educational field has attempted to improve teaching and teacher education, national and other government-based assessments can carry the risk that their results will be used primarily to rank educational systems or schools (Auld and Morris 2016). Equity issues such as gender differences in mathematics, or the achievement gap between majority and minority students, are also strongly aligned to the consequential aspects of validity, especially when assessments should inform measures to reduce inequality or are used to identify at-risk students. Those who analyse assessment data from computer-based assessments or learning analytics should be aware of the temporality of data and that today's assessment results will be yesterday's results tomorrow. Adding to the danger of neglecting this transitory nature are the simplifications and the implicit assumptions that both individual people and society as a whole make when response data are coded into algorithms. There is thus the risk that these kinds of assessment practices may reproduce and entrench existing biases such as class, gender, or ethnicity (Wilson et al. 2017).

A multitude of issues are connected to validity, such as the danger of using a single measurement point, the reliability of the outcome, or the validity of inferences made for different purposes. Pellegrino et al. (2001), for instance, claim that when "a single assessment is used for multiple purposes…the more purposes a single assessment aims to serve, the more each purpose will be compromised" (p. 2). Further, Newton (2007) points to the risks of ambiguous allocations of assessment purposes, since policy-makers could be misled if the complexities of the assessment design are over-simplified, for example when an assessment for a particular purpose (such as short-term system monitoring) is wrongly used for long-term system monitoring. Therefore, to scrutinise any potential consequences connected to the use of a particular assessment, different argument procedures might be applied, depending on whether the assessment was designed to assess individual students or for research or programme evaluation purposes at the institutional or national level.

6.3 Future directions for assessment in mathematics education

Numerous researchers within the research community deal with specific questions related to assessment in mathematics

education. In doing so, only isolated aspects of assessment development or implementation are taken into consideration, leaving other aspects in the dark. The strong relationships between teaching, learning, and assessment and between methodology, equity, and policy are challenging to disentangle without oversimplifying. For instance, ongoing revision of curricula, the further development of school practices, beliefs about the role of mathematics education, content, and teaching methods all contribute to further complexity. Correspondingly, assessment must be constantly adapted to new circumstances, such as increased heterogeneity in the classroom or new technological possibilities. We might see this as a daunting task, or it might encourage the research community to continue to work on the improvement of assessment practices in mathematics education. With regard to assessment practices in mathematics education, we see four key areas for future development, as described below.

1. *Efforts to develop more refined methods to improve the quality of educational assessment in mathematics education*. We propose that the high degree of complexity found in the assessment process should be taken into account to a larger extent than has previously been feasible. For instance, multiple assessment formats or multiple groups of test takers increase complexity and call for multiple approaches to validation. In the past, mixed-methods research has developed into an application-oriented research methodology, which is suitable for addressing validity questions when these questions are transferred to assessment practice. Applying mixed-methods research methods offers the possibility of managing this complexity in new ways.
2. *Efforts to improve the relation between research and practice*. We propose that the research community should emphasise to a larger extent the relevance of assessment research to teaching practice and educational policy. Many see educational research and teaching practice as different reference systems that coexist independently of each other and that have different orientations, which could explain why recent research results from international large-scale studies or from effectiveness research often have little practical significance or transferability (Burkhardt and Schoenfeld 2003, 2018). By taking into account the increased heterogeneity found in classrooms and student backgrounds when analysing assessment data, those in the education assessment field will be able to provide more applicable advice for practice and policy. By doing so, we can increase the scope by which the respective stakeholders involved can view the outcomes of both assessment and research studies as being valid, thus contributing to the quality of mathematics education.
3. *Efforts to improve equity*. We propose that culturally responsive assessment represents a decisive further development in the area of addressing heterogeneity in assessment that takes into account the relationship between equity issues and educational policy. We must extend the notion of cultural responsiveness by investigating how large-scale assessment in mathematics can allow students from diverse cultural backgrounds to participate. Further investigations might include new approaches to researching the influence of language on mathematics achievement in assessment situations, or the use of formative assessment in classrooms and teacher education.
4. *Applying technology to develop better measures of mathematical competence*. The further development of computers, software, and digital tools has pushed forward the question of whether, how, and to what extent we can use technology to assess mathematical knowledge, thinking, or skills. Recent developments have strongly influenced assessment practice; we propose that those in the field should use this knowledge to develop better measures of students' mathematical competence, while taking into consideration not only the multiple possibilities but also the technical and methodological challenges involved.

In reality, we cannot look at these four key areas for future development independently. In order to work towards equity, we might utilise technology to develop assessments where language is understood more broadly; such assessments might incorporate animations and visual displays, for example. Applying mixed methods to learn more from today's assessment formats and practices is a key to further developing better tests and practices that can provide teachers and policy-makers with the insights they need to improve their practices. Further, we might utilise a wider range of assessment formats to enable both equity in mathematics education and more suitable assessment data for policy-making. Finally, educational policies might affect what is assessed and how; thus, such policies may contribute to more equitable practices in mathematics education. A stronger link between research and teaching might be facilitated by carefully considering the issues related to methodology, policy, and equity discussed in this paper; such considerations should concern each issue individually as well as the relationships between the three issues and how they influence one another and assessment validity. Because we use assessment outcomes to inform teaching, select students for further education, certify professionals, and shape educational policies, it is vital that we discuss the technical affordances and possibilities that each assessment format offers. Clearly, a special issue on assessment in mathematics education can

help us to address important challenges in our field of scientific inquiry.

References

Abedi, J., & Lord, C. (2001). The language factors in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234.

Auld, E., & Morris, P. (2016). PISA, policy and persuasion: Translating complex conditions into education 'best practice'. *Comparative Education, 52*(2), 202–229.

Australian Association of Mathematics Teachers Inc. (2008). Position paper on the practice of assessing mathematical learning. http://www.aamt.edu.au/content/download/9895/126744/file/Assessment_position_paper_2017.pdf. Accessed 9 July 2017.

Ayalon, H., & Livneh, I. (2013). Educational standardization and gender differences in mathematics achievement: A comparative study. *Social Science Research, 42*(2), 432–445.

Baird, J.-A., Johnson, S., Hopfenbeck, T. H., Isaacs, T., Sprague, T., Stobart, G., & Yu, G. (2016). On the supranational spell of PISA in policy. *Educational Research, 58*(2), 121–138.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., et al. (2010). Problems with the use of student test scores to evaluate teachers. Economic Policy Institute Briefing Paper #278. http://www.epi.org/publication/bp278/. Accessed 9 July 2017.

Biesta, G. (2009). Good education in an age of measurement: On the need to reconnect with the question of purpose in education. *Educational Assessment, Evaluation and Accountability, 21*(1), 33–46.

Black, P., & Wiliam, D. (2005). Inside the black box: Raising standards through classroom assessment. *The Phi Delta Kappan, 80*(2), 139–148.

Black, P., & Wiliam, D. (2012). Assessment for learning in the classroom. In J. Gardner (Ed.), *Assessment and learning* (pp. 11–32). London: Sage.

Bradshaw, C. P., O'Brennan, L. M., & McNeely, C. A. (2008). Core competencies and the prevention of school failure and early school leaving. *New Directions for Child and Adolescent Development, 122*, 19–32.

Brown, G. T. L., & Harris, L. R. (2009). Unintended consequences of using tests to improve learning: How improvement-oriented resources heighten conceptions of assessment as school accountability. *Journal of Multidisciplinary Evaluation, 6*(12), 68–91.

Buchholtz, N., Kaiser, G., & Blömeke, S. (2014). Measuring pedagogical content knowledge in mathematics—conceptualizing a complex domain. *Journal für Mathematik-Didaktik, 35*(1), 101–128.

Buchholtz, N., Krosanke, N., Orschulik, A. B., & Vorhölter, K. (2018). Combining and integrating formative and summative assessment in mathematics teacher education. *ZDM Mathematics Education, 50*(4), 1–14.

Buchholtz, N., Leung, F. K. S., Ding, L., Kaiser, G., Park, K., & Schwarz, B. (2013). Future mathematics teachers' professional knowledge of elementary mathematics from an advanced standpoint. *ZDM, 45*(1), 107–120.

Burkhardt, H., & Schoenfeld, A. (2003). Improving educational research: Toward a more useful, more influential, and better-funded enterprise. *Educational Researcher, 32*(9), 3–14.

Burkhardt, H., & Schoenfeld, A. (2018). Assessment in the service of learning: Challenges and opportunities. *ZDM Mathematics Education, 50*(4), 1–15.

Cai, J., Hwang, S., & Middleton, J. A. (2015). The role of large-scale studies in mathematics education. In J. A. Middleton, S. Hwang & J. Cai (Eds.), *Large-scale studies in mathematics education* (pp. 405–414). Cham: Springer.

Cai, J., Mok, I. A. C., Reddy, V., & Stacey, K. (2016). International comparative studies in mathematics: Lessons for improving students learning. In *ICME-13 topical surveys* (pp. 1–36). Cham (Switzerland): Springer.

Cotton, C., McIntyre, F., & Price, J. (2010). Gender differences disappear with exposure to competition. Working paper 2010–11. University of Miami, Department of Economics. http://moya.bus.miami.edu/~ccotton/papers/cotton_mcintyre_price_2009.pdf. Accessed 9 July 2017.

Elstad, E., Nortvedt, G. A., & Turmo, A. (2009). The Norwegian assessment system: An accountability perspective. *CADMO, 17*(1), 89–103.

Ernest, P. (2014). Policy debates in mathematics education. In S. Lerman (Ed.), *Encyclopedia of mathematics education*. Dordrecht: Springer.

Fischer, R. (2004). Standardization to account for cross-cultural response bias: A classification of score adjustment procedures and review of research. *Journal of Cross-Cultural Psychology, 35*(3), 263–282.

Fujita, T., Jones, K., & Miyazaki, M. (2018). Learners' use of domain-specific computer-based feedback to overcome logical circularity in deductive proving in geometry. *ZDM Mathematics Education, 50*(4), 1–15.

Gaber, S., Cankar, G., Umek, L. M., & Tašner, V. (2012). The danger of inadequate conceptualisation in PISA for education policy. *Compare, 42*(4), 647–663.

Grant, M., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal, 26*(2), 91–108.

Groß Ophoff, J. (2013). *Lernstandserhebungen: Reflexion und Nutzung*. Münster: Waxmann.

Hallinger, P., & Heck, R. H. (2010). Collaborative leadership and school improvement: Understanding the impact on school capacity and student learning. *School Leadership & Management, 30*(2), 95–110.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J. L., et al. (2007). *Standards-based accountability under no child left behind: Experiences of teachers and administrators in three states*. Santa Monica: RAND Corporation.

Hannon, B. (2012). Test anxiety and perfomance-avoidance goals explain gender differences in SAT-V, SAT-M, and overall SAT scores. *Personality and Individual Differences, 53*(7), 816–820.

Hattie, J. A. C., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Heritage, M., & Wylie, C. (2018). Reaping the benefits of assessment for learning: Achievement, identity and equity. *ZDM Mathematics Education, 50*(4), 1–13.

Hoogland, K., & Tout, D. (2018). Computer-based assessment of mathematics in the 21st century: Pressures and tensions. *ZDM Mathematics Education, 50*(4), 1–12.

Hopfenbeck, T. H., & Görgen, K. (2017). The politics of PISA: The media, policy and public responses in Norway and England. *European Journal of Education, 52*(2), 195–205.

Hopson, R., & Hood, S. (2005). An untold story in evaluation roots: Reid E. Jackson and his contribution toward culturally responsive evaluation at three quarters of a century. In S. Hood, R. Hopson

& H. Frierson (Eds.), *The role of culture and cultural context* (pp. 87–104). Greenwich: Information Age Publishing.

Hoth, J., Döhrmann, M., Kaiser, G., Busse, A., König, J., & Blömeke, S. (2016). Diagnostic competence of primary school mathematics teachers during classroom situations.. *ZDM Mathematics Education, 48*(1), 41–53.

Hsieh, F.-J., Chu, C.-T., Hsieh, C.-J., & Lin, P.-J. (2014). In-depth analyses of different countries' responses to MCK items: A view on the differences within and between East and West. In S. Blömeke, F.-J. Hsieh, G. Kaiser & W. H. Schmidt (Eds.), *International perspectives on teacher knowledge, beliefs and opportunities to learn* (pp. 115–140). Dordrecht: Springer.

Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences of the United States of America, 106*(22), 8801–8807.

Institut zur Qualitätsentwicklung im Bildungswesen (IQB). (2017). *Erprobungsstudie 2017 zu den Bildungsstandards Mathematik in der Sekundarstufe* I. https://www.iqb.hu-berlin.de/bt/BT2018/Erprobungsstudie2017. Accessed 27 Apr 2018.

Jerrim, J. (2016). PISA 2012: How do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practice, 23*(4), 495–518.

Kaarstein, H. (2014). Norwegian mathematics teachers' and educational researchers' perception of MPCK items used in the TEDS-M study. *Nordisk Matematikkdidaktikk, 19*(3–4), 57–82.

Kaiser, G., Blömeke, S., König, J., Busse, A., Döhrmann, M., & Hoth, J. (2017). Professional competencies of (prospective) mathematics teachers: Cognitive versus situated approaches. *Educational Studies in Mathematics, 94*(2), 161–182.

Kilpatrick, J. (2014). History of research in mathematics education. In S. Lerman (Ed.), *Encyclopedia of mathematics education*. Dordrecht: Springer.

Klenowski, V. (2009). Australian indigenous students: Addressing equity issues in assessment. *Teacher Education, 20*(1), 77–93.

Leder, G., & Forgasz, H. J. (2018). Measuring who counts: Gender and mathematics assessment. *ZDM Mathematics Education*, *50*(4), 1–11.

Lester, F. Jr. (Ed.). (2007). *Second handbook of research on mathematics teaching and learning*. Charlotte: Information Age Publishing.

Lin, F.-L., Wang, T.-Y., & Chang, Y.-P. (2018). Effects of large-scale studies on mathematics education policy on Taiwan through the lens of societal and cultural characteristics. *ZDM Mathematics Education*, *50*(4), 1–14.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*(6), 1123–1135.

Liu, O. L., & Wilson, M. (2009). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. *Applied Measurement in Education, 22*(2), 164–184.

Lynch, K., & Star, J. R. (2014). Teachers' views about multiple strategies in middle and high school mathematics. *Mathematical Thinking and Learning, 16*(2), 85–108.

Ma, X. (1999). A meta-analysis of the relationship between anxiety towards mathematics and achievement in mathematics. *Journal for Research in Mathematics Education, 30*(5), 520–540.

Martinovic, D., & Manizade, A. G. (2018). The challenges in the assessment for knowledge for teaching geometry. *ZDM Mathematics Education*, *50*(4), 1–17.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Middleton, J. A., Cai, J., & Hwang, S. (2015). Why mathematics education needs large-scale research. In J. A. Middleton, J. Cai & S. Hwang (Eds.), *Large-scale studies in mathematics education* (pp. 1–3). Cham: Springer.

Miller, J., & Mitchell, J. (2006). Interrupted schooling and the acquisition of literacy: Experiences of Sudanese refugees in Victorian secondary schools. *Australian Journal of Language and Literacy, 29*(2), 150–162.

Montenegro, E., & Jankowski, N. A. (2017). Equity and assessment: Moving towards culturally responsive assessment. National Institute for Learning Outcomes Assessment. http://learningoutcomesassessment.org/documents/OccasionalPaper29.pdf. Accessed 9. July 2017.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). TIMSS 2015 international results in mathematics. Boston College TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/international-results/. Accessed 9 July 2017.

Museus, S. D., Palmer, R. T., Davis, R. J., & Maramba, D. (2011). Special issue: Racial and ethnic minority student success in STEM education. *ASHE Higher Education Report, 36*, 1–140.

National Council of Teachers of Mathematics (NCTM). (2016). *Large-scale mathematics assessments and high-stakes decisions: A position of the National Council of Teachers of Mathematics*. Reston: NCTM.

National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Neubrand, M. (2018). Conceptualizations of professional knowledge for teachers of mathematics. *ZDM Mathematics Education*, *50*(4), 1–12.

Newton, P. E. (2007). Clarifying the purpose of educational assessment. *Assessment in Education: Principles, Policy & Practice, 14*(2), 149–170.

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational and psychological assessment*. London: Sage.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge: Harvard Education Press.

Niss, M. (1993). Assessment in mathematics education and its effects: An Introduction. In M. Niss (Ed.), *Investigations into assessment in mathematics education. An ICMI Study* (pp. 1–30). Dordrecht: Springer.

Niss, M. (2007). Reflections on the state of and trends in research on mathematics teaching and learning. In F. K. J. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 1293–1312). Charlotte: Information Age Publishing.

Niss, M. (2015). Mathematical competencies and PISA. In R. Turner & K. Stacey (Eds.), *Assessing mathematical literacy: The PISA experience* (pp. 35–55). Cham: Springer.

Nortvedt, G. A. (2011). Coping strategies applied to comprehend multistep arithmetic word problems by students with above-average numeracy skills and below-average reading skills. *Journal for Mathematical Behavior, 30*(3), 255–269.

Nortvedt, G. A. (2018). Policy impact of PISA on mathematics education: The case of Norway. *European Journal for Psychology in Education, 33*(3), 427–444.

Nortvedt, G. A., Gustafsson, J.-E., & Lehre, A.-C. W. G. (2016). The importance of InQua for the relation between achievement in reading and mathematics. In T. Nilsen & J.-E. Gustafsson (Eds.), *Teacher quality, instructional quality and student outcome: Relationships across countries, cohorts and time* (pp. 97–113). Cham: Springer.

OECD. (2013a). *PISA 2012 results: Student performance in mathematics, reading, science. Volume I*. Paris: OECD Publishing.

OECD. (2013b). *PISA 2012 results: Ready to learn. Students' engagement, drive and self-beliefs. Volume III*. Paris: OECD Publishing.

OECD. (2013c). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris: OECD Publishing.
OECD. (2015). *Helping immigrant students to succeed at school—and beyond*. Paris: OECD Publishing.
OECD. (2016). *PISA 2015 results: Excellence and equity in education* (Vol I). Paris: OECD Publishing.
Pajares, F., & Miller, M. D. (1995). Mathematics self-efficacy and mathematics performances: The need for specificity of assessment. *Journal of Counseling Psychology, 42*(2), 190–198.
Palm, T., Boesen, J., & Lithner, J. (2011). Mathematical reasoning in Swedish upper secondary level assessments. *Mathematics Thinking and Learning, 13*(3), 221–246.
Pankow, L., Kaiser, G., & König, J. (2018). Perception of students' errors under time limitation: Are teachers better than mathematicians or students? Results of a validation study. *ZDM Mathematics Education*, *50*(4), 1–12.
Paxton, G., Smith, N., Win, A. K., Mulholland, N., & Hood, S. (2011). *Refugee status report: A report on how refugee children and young people in Victoria are faring*. Melbourne: Department of Education and Early Childhood Development (DEECD).
Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
Rowland, T., & Ruthven, K. (2010). *Mathematical knowledge in teaching*. Dordrecht: Springer.
Sälzer, C., & Prenzel, M. (2014). Looking back at five rounds of PISA: Impacts on teaching and learning in Germany. *Solsko Polje, 25*(5/6), 53–72.
Sangwin, C. J. (2013). *Computer aided assessment of mathematics*. Oxford: Oxford University Press.
Semana, S., & Santos, L. (2018). Self-regulation of learning in student participation in mathematics assessment. *ZDM Mathematics Education*, *50*(4), 1–13.
Scherer, P., Beswick, K., DeBois, L., Healy, L., & Opitz, E. M. (2016). Assistance of students with mathematical learning difficulties: How can research support practice? *ZDM, 48*, 633–649.
Schoenfeld, A. (2007). Issues and tensions in the assessment of mathematical proficiency. In A. Schoenfeld (Ed.), *Assessing mathematical proficiency* (pp. 3–16). New York: Cambridge University Press.
Seeley, C. (2006). Teaching to the test. *NCTM News Bulletin*. http://www.nctm.org/News-and-Calendar/Messages-from-the-President/Archive/Cathy-Seeley/Teaching-to-the-Test/. Accessed 9 July 2017.
Shen, C., & Tam, H. P. (2008). The paradoxical relationship between student achievement and self-perception: A cross-national analysis based on three waves of TIMSS data. *Educational Research and Evaluation, 14*(1), 87–100.
Siemon, D., Enilane, F., & McCarty, J. (2004). Supporting indigenous students' achievement in numeracy. *Australian Primary Mathematics Classroom, 9*(4), 50–53.
Speer, N. M., King, K. D., & Howell, H. (2015). Definitions of mathematical knowledge for teaching: Using these constructs in research on secondary and college mathematics teachers. *Journal of Mathematics Teacher Education, 18*(2), 105–122.
Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. Oxford: Routledge.
Suurtamm, C., & Neubrand, M. (2015). Assessment and testing in mathematics education. In S. J. Cho (Ed.), *Proceedings of the 12th International Congress on Mathematical Education* (pp. 557–562). Cham: Springer.
Suurtamm, C., Thompson, D. R., Kim, R. Y., Moreno, L. D., Sayac, N., Schukajlow, S., et al. (2016). *Assessment in mathematics education: Large-scale assessment and classroom assessment*. Cham: Springer.
Ubuz, B., Aydin. (2018). Geometry knowledge test about triangles: Development and validation. *ZDM Mathematics Education, 50*(4).
van den Heuvel-Panhuizen, M., & Becker, J. (2003). Towards a didactic model for assessment design in mathematics education. In A. J. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick & F. K. S. Leung (Eds.), *Second international handbook of mathematics education* (pp. 689–716). Dordrecht: Springer.
Wang, S., Jiao, H., Young, M., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K–12 mathematics tests. *Educational and Psychological Measurement, 67*(2), 219–238.
Wiliam, D. (2003). The impact of educational research on mathematics education. In A. J. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick & F. K. S. Leung (Eds.), *Second international handbook of mathematics education* (pp. 471–490). Dordrecht: Springer Netherlands.
Wiliam, D. (2007). Keeping learning on track. In F. K. J. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 1053–1098). Charlotte: Information Age.
Wilson, A., Watson, C., Thompson, T. L., Drew, V., & Doyle, S. (2017). Learning analytics: Challenges and limitations. *Teaching in Higher Education, 22*(8), 991–1007.
Wong, P. A., & Glass, R. D. (2005). Assessing a professional development school approach to preparing teachers for urban schools serving low-income, culturally and linguistically diverse communities. *Teacher Education Quarterly, 32*(3), 63–77.
Wößmann, L. (2005). The effect heterogeneity of central examinations: Evidence from TIMSS, TIMSS-Repeat and PISA. *Education Economics, 13*(2), 143–169.
Wuttke, J. (2007). Uncertainties and bias in PISA. In S. T. Hopmann, G. Brinek & M. Retzl (Eds.), *PISA according to PISA: Does PISA keep what it promises?* Vienna: LIT-Verlag.
Hansen, K. Y., & Strietholt, R. (2018). Does schooling actually perpetuate educational inequality in mathematics performance? A question of validity. *ZDM Mathematics Education*, *50*(4), 1–6.