International Journal on Social and Education Sciences
Volume 2, Issue 1, 2020 ISSN: 2688-7061 (Online)

Developing a Validated Test to Measure Students' Progression in Mathematical Reasoning in Primary School

Dorte Moeskær Larsen
Southern University of Denmark, Denmark, dmla@imada.sdu.dk

Morten Rasmus Puck
UCL University College, Denmark

Abstract: Not enough effort has been invested in developing reliable and valid assessment instruments to measure students' development of reasoning competences in mathematics. Previously developed tests rely mostly on standardized multiple-choice assessments, which primarily focus on procedural knowledge and rote learning and not on how students argue for and justify their results. This study reports on the process of developing a test to assess students' reasoning competence in primary school classes and on the statistical results of a verification study. The test is based on item response theory (Rasch model) and was used in a large-scale mathematics intervention project about inquiry-based teaching. The data was collected from 5,516 primary school students in Denmark. Data analysis indicates the test has satisfactory validity and reliability, but the test was still unable to measure statistically significant progress at the intervention schools.

Keywords: Assessment, Item Response Theory, Mathematical reasoning, Primary school

Introduction

Despite the importance of reasoning in school mathematics (Ball & Bass, 2003; Hanna & Jahnke, 1996), research shows that many students at all school levels face serious difficulties with reasoning and proving (EMS, 2011; Harel & Sowder, 1998). Consequently, children enter upper grades ill-equipped to develop their justifications and proofs (Knuth, 2002). Multiple factors may have contributed to the fact that reasoning and proof have received a marginal role in primary and lower secondary school mathematics teaching (G. J. Stylianides & Stylianides, 2017).

However, one particular hindrance could be that, if the way we assess students is focused only on procedural knowledge and not on reasoning competences, there is a risk that tests can obstruct the development of more complex mathematical competences by distorting the focus in teaching away from a competence orientation. As Biggs (2011, p. 197) argued, assessment determines what and how students learn more than a curriculum does. Even in achievement tests in which the primary recipients are people at some distance who require an assessment of an overall effect, rather than detailed information on individual students, assessments can have a powerful influence on teaching and learning: 'we simply want to stress that accountability tests, by virtue of their place in a complex social system, exercise an important influence on the curriculum of the school' (Resnick & Resnick, 1992, p. 49). As a consequence, achievement tests need to be closely aligned with the target knowledge of the curriculum, but, although reasoning is well-known in mathematical teaching practice and has been the focal point of many studies (Lithner, 2008; Reid & Knipping, 2010; A. J. Stylianides & Harel, 2018), there is limited research on how to assess students' competence with reasoning in mathematics education.

This paper introduces a newly developed large-scale achievement test that uses a Rasch model and is intended to measure students' development of competence in mathematics. To determine the best way of conducting assessment in a particular case and context, it is necessary to consider the properties of possible tools in relation to the purposes and intended uses of the assessment results. Obvious important desirable properties of any assessment are its reliability and its validity for the intended purpose (Johnson & Christensen, 2014), which will be the focus in this paper.

Theoretical Background

Developing a test that can measure reasoning competence requires both a study of how others have tried to measure it in the past and consideration of how to make the test valid and reliable. In the following, these two considerations will be elaborated.

Testing Reasoning in Mathematics Education

In both of the two major international large-scale assessments – the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA) – we have seen some effort to measure reasoning competence, but, in general, the testing of competences in mathematics education is not a widespread nor well-documented area. Niss, Bruder, Planas, Turner, and Villa-Ochoa (2016) argued that assessment of students' mathematical competences needs to become a priority in educational research from both holistic and atomistic perspectives, where a holistic perspective considers complexes of intertwined competences in the enactment of mathematics and 'an atomistic perspective zooms in on the assessment of the individual competency in contexts stripped, as much as possible, of the presence of other competences' (Niss et al., 2016, p. 624).

Through a literature review of assessment in mathematics (Larsen, 2017), created with a focus on developing such a test, different aspects of measuring reasoning in mathematics were found. Logan and Lowrie (2013) focused on testing students' spatial reasoning in mathematics. In Nunes, Bryant, Evans, and Barros (2015), a framework for prescribing and assessing the inductive mathematics reasoning of primary school students was formulated and validated. The major constructs incorporated in this framework were students' cognitive abilities for finding similarities and/or dissimilarities among the attributes and relationships of mathematical concepts. Nunes et al. (2007) focused on reasoning in logic, while Goetz, Preckel, Pekrun, and Hall (2007) examined reasoning in connection to another aspect – test-related experiences of enjoyment, anger, anxiety, and boredom and how they relate to students' abstract reasoning ability.

Overall, we see that the term reasoning is broadly used in mathematics education and is connected to many different areas, such as logic, spatial, and cognitive abilities. Reasoning in mathematics is therefore not a unified concept, which makes an assessment of this competence a challenge. Yackel and Hanna (2003) argued that the reason for disagreement on the definition is, in fact, an implicit assumption of universal agreement on its meaning, yet there are many different conceptualizations of mathematical reasoning in the literature.

Brousseau and Gibel (2005) defined mathematical reasoning as a relationship between two elements: a condition or observed facts and a consequence. Duval (2007) described mathematical reasoning as 'a logical linking of propositions' (p. 140) that may change the epistemic value of a claim. G. J. Stylianides (2008) viewed mathematical reasoning as containing a broader concept; besides providing arguments (non-proof and proofs), it also encompasses investigating patterns and making conjectures. In Shaughnessy, Barrett, Billstein, Kranendonk, and Peck (2004), the National Council of Teachers of Mathematics (NCTM) described reasoning in mathematics as a cycle of exploration, conjecture, and justification. This is in line with G. J. Stylianides (2008), who focused on mathematical reasoning as a process of making inquiries to formulate a generalisation or conjecture and determine its truth value by developing arguments, where argument is understood to mean a connected sequence of assertions. In a Danish context, the definition of reasoning competence in the primary school mathematics curriculum is often based on the competence report (Niss & Jensen, 2002), which, among other factors, is very specific about students being able to distinguish proof from other kinds of argumentation and makes a clear distinction between 'carrying out' argumentations and 'following' the argumentations developed by others (e.g., other students or textbooks). In the present paper, the definition of reasoning that will be tested in the developed testis a composite of definitions from the NCTM, the competence report (Niss & Jensen, 2002), and G. J. Stylianides (2008). This will be described in more detail in the Methods section.

Reliability and Validity in Tests

The term validity refers to whether or not a test measures what it claims to measure. On a test with high validity, the items will be closely linked to the test's intended purpose. In 1955, Cronbach and Meehl wrote the classic article, Construct Validity in Psychological Tests, in which they divided test validity into three types: content, construct, and criterion-oriented. For many years, the discussion has been about these different types of validity, but, in this paper, the validity issue will be focused more on obtaining evidence for unitary validity (Johnson &

Christensen, 2014), which includes all three types of validity but in which the central question is whether all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose.

The intention in this study is therefore to present and discuss whether the collected evidence supports the argument that the test can be seen as valid. Johnson and Christensen (2014) argued that, in this sense, complete validation is never fully attained – it is not a question of no validity versus complete validity. Validation should be viewed as a never-ending process, but, at the same time, the more validity evidence one has, the more confidence one can place in one's interpretations.

The other aim in the development of this test is for it to provide reliable or trustworthy data. If a test provides reliable scores, the scores will be similar on every occasion, which is related to generalization from the results (Black, 1998). In summary, the aims of this study are to develop a test that, among other things, can measure reasoning competence in primary school classes and to examine the quality of that test.

Methods

The methods section starts with at short description of the setting of the assessment. A description then follows of how the development process was divided into three different phases: design, development, and testing. The three phases are briefly reviewed, followed by information on how they relate to each other. The text then describes phase 1, which, in addition to a description of the definition of reasoning, also includes an example of an item from the test. In phase 2, the pilot study and the measurement model are described, including intercoder reliability, and, in phase 3, elaboration and consideration of content validity are described.

The Origin of and Reasons for Developing a Competence Test

This study is embedded in a large-scale, three-year, design-based, randomized-controlled-trial research and development programme in Denmark, called 'Kvalitet i Dansk og Matematik' (Quality in Danish and Mathematics; KiDM). The overall aim of KiDM is to make teaching in years 4 and 5 of Danish primary school more inquiry-based, and it includes the development of inquiry-based teaching activities for a four-month mathematics teaching plan implemented in 45 Danish schools (intervention schools). To assess students' development of mathematical competences related to inquiry-based teaching, an achievement test, with a strong focus on mathematical competences, was essential to measure the difference between the 45 interventions schools and 37 control schools. A test was therefore developed specifically to measure whether students developed specific competences in mathematics. These competences included those of problem-solving/-posing, modeling, and reasoning.

In this study, we will focus on whether this test is valid and reliable in connection to the part of the test connected to reasoning competence. The intent in the test is to use a Rasch model analysis, which is a psychometric technique that can improve the precision with which researchers construct instruments, monitor instrument quality, and compute respondents' performances (Wilson & Gochyyev, 2013). Rasch analysis allows researchers to construct alternative forms of measurement instrument because it does not treat a test scale as linear, with the raw scores of different respondents simply 'added up' to compare their levels of achievement. Rasch analysis instead allows researchers to use respondents' scores to express their performance on a scale that accounts for the unequal difficulties across the test items. It is also a tool for finding items that measure the same phenomenon.

Designing, Developing, and Testing the Test

In this section, the development of the item design and the coding processes will be elaborated, together with the verification. The development was done in an iterative process comprising three different phases before the test was finally applied in three randomized controlled trials as part of KiDM.

> Phase 1: Designing the test: The components in the test (content and design) were based on theory identified in the established literature in collaboration between the authors of this paper, two associate professors from two different mathematics teacher education colleges, and one professor in mathematics education.

Phase 2: Developing the test: Teachers from a primary school class and their supervisor in mathematics tested the test in their classroom, and interviews with three of their students were conducted. Think-alouds were also conducted with one student from year 5 and one student from year 4.
Phase 3: Testing the test: The test was piloted in 14 classes, with a focus on items appropriate to the intended pupils; the coding procedures were further developed, and the test administration was itself tested.
The three phases were not neatly divided because the various tests and trials required some corrections to the test, including the difficulty of the items, and the development of the coding process during the pilot test required changes to the content and the formulations, which needed consideration of the design and content, as will be described later.

Phase 1: Designing the test. In order for the test to measure the content of reasoning competence, which is aligned with the definition of reasoning competence in KiDM, it was decided that the definition should be process-oriented but also have a broad approach. Therefore, definitions from the NCTM, the competence report (Niss & Jensen, 2002), and G. J. Stylianides (2008) were all included. In Table 1, the three different broad definitions from the Introduction section are included in the left column to allow comparison of the definitions, and, in the right column is listed what is intended to be measured by different items in the test (only the final included items are listed).

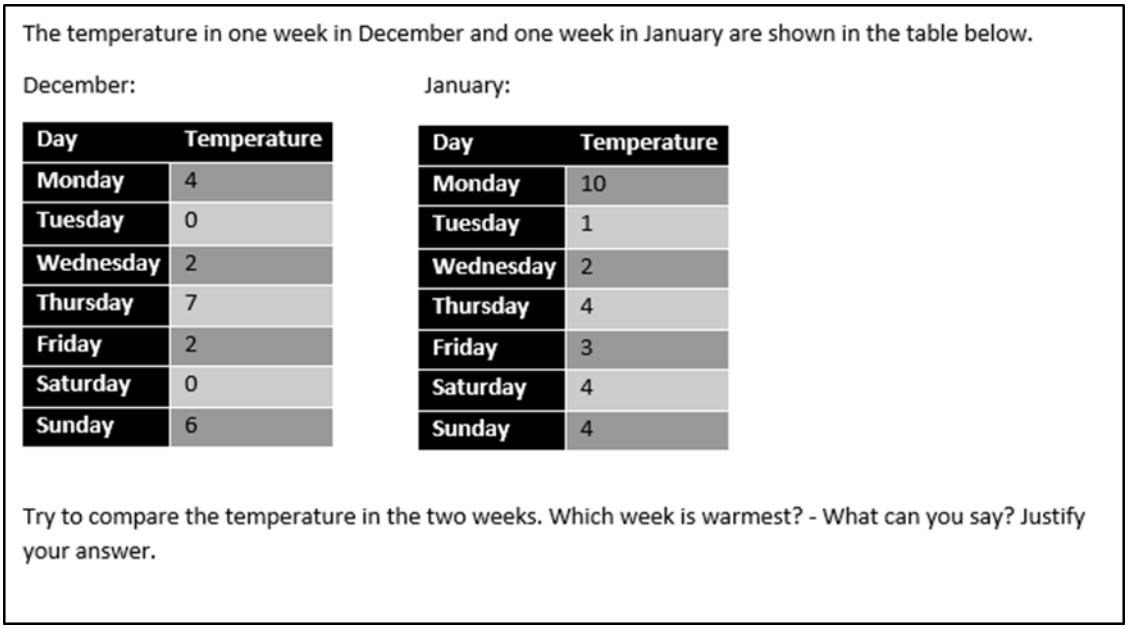
Table 1. Content of Items Included in the Final KiDM Test

	Standards from NCTM	**Niss & Jensen (2002)**	**G. J. Stylianides (2008)**	**The KiDM test**	**Items with this focus (item number)***
A	Make and investigate mathematical conjectures		Making conjectures	**Can the students make inquiries to formulate assumptions/conjectures in mathematics?**	**1, 9, 10, 14, 15, 17, 18, 19, 26**
B			Investigating patterns and making generalizations	**Can the students make inquiries to formulate generalizations in mathematics?**	**13, 18**
C	Develop mathematical arguments and proofs	To devise informal and formal reasoning (on the basis of intuition), including transforming heuristic reasoning into actual (valid) proof	Providing support to mathematical claims; providing non-proof argumentations and providing proofs	**Can the students develop argumentations in mathematics? (proof/non-proof)**	**1, 9, 17**
D	Evaluate mathematical arguments and proofs	To understand and judge a mathematical argumentation propounded by others		**Can the students understand and judge argumentations in mathematics? (proof/non-proof)**	**15, 19, 29**
E	Use various types of reasoning and methods of proof; select and use various types of reasoning and methods of proof	To devise and implement informal and formal reasoning (on the basis of intuition), including transforming heuristic reasoning into actual (valid) proof	Providing support to mathematical claims; providing non-proof argumentations and providing proofs	**Can the students select and use various types of reasoning?**	**28, 29**
F	*Select and use various types of reasoning and methods of proof*	*To know and understand what a mathematical proof is and how it is different from other argumentations*		***Can the students distinguish between different kinds of arguments – rationales, empirical arguments, and deductive arguments?***	***Items with this content have been removed because they were found too difficult for the students in years 4 and 5***

*Note: the remaining items focus on modeling competence or problem-solving/-posing competence.

The items are the substance of the test and have been developed with inspiration both from the research, such as A. J. Stylianides (2007), and from published TIMSS and PISA items as both tests explicitly state that they measure mathematical reasoning and, moreover, TIMSS is aimed at the same age range as our population. The items included in the test are a mixture of multiple-choice and closed-constructed response questions as well as open-constructed problems that include, for example, making argumentation for one's own conjecture or justifying one's choice of which argumentation is correct in different problems. This specific mix was chosen because we wanted to use different types of item to accommodate different approaches to those items, but also because open items are very costly in the coding process and, therefore, it was not possible to have only open-constructed items. Multiple-choice items and closed-constructed items, in contrast, can be automatically scored by computer and are, therefore, almost cost-free.

The open-constructed items were, in the KiDM project, manually scored by three (two in the last trial) preservice teachers. The open-constructed items include a mixture of dichotomous and polytomous questions; the dichotomous questions can only be coded 0 or 1 (incorrect or correct), while the polytomous questions have up to four different scoring categories. In Figure 1, an open-constructed item (item 9) from the KiDM test is presented that explores the similarities and differences between relationships in statistics in a close-to-reality task. In this task, the student must show, in connection to A and C in Table 1, that he or she is able to make inquiries to formulate a conjecture and develop argumentation to verify his or her conjecture.



The temperature in one week in December and one week in January are shown in the table below.

December:

Day	Temperature
Monday	4
Tuesday	0
Wednesday	2
Thursday	7
Friday	2
Saturday	0
Sunday	6

January:

Day	Temperature
Monday	10
Tuesday	1
Wednesday	2
Thursday	4
Friday	3
Saturday	4
Sunday	4

Try to compare the temperature in the two weeks. Which week is warmest? - What can you say? Justify your answer.

Figure 1. Test Item 9: Comparing Temperatures in Two Different Weeks

In order to differentiate the quality of students' answers to the different open-ended items, coding taxonomies were developed for each. These coding taxonomies (schemes) were based on a theoretical progression described in the literature; however, the concept of argumentation has very different definitions in the literature, and, to explain item 9's coding scheme, we first look at theoretical descriptions of different approaches to taxonomies.
Arguments are often divided into non-proof arguments and deductive proof arguments (Reid & Knipping, 2010). A deductive argument is one that, if valid, has a conclusion that is entailed by its premises. In other words, the truth of the conclusion is a logical consequence of the premises and, if the premises are true, the conclusion must also be true. It would be self-contradictory to assert the premises and deny the conclusion, because the negation of the conclusion is contradictory to the truth of the premises. Therefore, deductive arguments may be either valid or invalid (Reid & Knipping, 2010).

G. J. Stylianides (2008) distinguished between two kinds of non-proof argument: empirical arguments and rationales. An argument counts as a rationale if it does not make explicit reference to some key accepted truths that it relies on or if it relies on statements that do not belong to the set of accepted truths of a community. Harel and Sowder (1998) introduced the concept of proof schemes, which classify what constitutes ascertaining and persuading for a person (or community). The taxonomy of proof schemes consists of three classes: external conviction, empirical conviction, and deductive conviction. Within the external conviction class of proof schemes, to prove depends on an authority, such as a teacher or a book. It could also be based strictly on the

appearance of the arguments or on symbol manipulations. Proving, within the empirical proof scheme class, is marked by reliance on either evidence from examples or direct measurement of quantities, substitutions of specific numbers, or perceptions. The third, deductive proof scheme class consists of two subcategories, each consistent with different proof schemes: the transformational proof scheme and the axiomatic proof scheme.
In Table 2, the created theoretical taxonomy of items about 'inquiry to make a conjecture' and 'developing argumentation' (such as item 9 in Figure 1) is presented, based on theories from Harel and Sowder (1998) and G. J. Stylianides (2008). The scale is ordinal.

At Location I, students are only able to develop a conjecture, and there are no arguments at all. At Location II, the conjecture is explicit, but the argument is only a rationale (G. J. Stylianides, 2008) or from an external conviction proof scheme (Harel & Sowder, 1998). Location III includes students who can support a developed conjecture with an empirical argument. Locations IV and V indicate students who can make a conjecture, draw implications by linking pieces of information from aspects of the problem, and make arguments from one side (Location IV) or more than one side (Location V). This resembles what Harel and Sowder (1998) called a deductive proof scheme. In the design phase, the research group's main focus was on whether each textual passage and its associated items adhered to a theoretical model, but, in the development and testing phases, the main focus was on whether the students' responses were associated with these theories.

Table 2. A Theoretical Model of the Taxonomy of Reasoning in Mathematics for Item 9

Location	Taxonomy developed from theory	Coding scheme for item 9	Codes for item 9
V	Develop/follow/critique assumptions/claims/statements with two-sided comparative deductive/mathematical arguments. Draw implications by linking pieces of information from separate aspects of the problem and make a chain of argumentation using deduction/mathematical concepts.		
IV	Develop/follow/critique assumptions/claims/statements supported with one-sided mathematical/deductive arguments. Draw implications by linking information from the problem with argument(s).	'The mean is a good way of comparing the temperature, and the mean is higher in January than in December because, in December, the mean is 3 and in January it is 4. We can say that it is warmer in January.'	3
III	Develop/follow/critique assumptions/claims/statements supported with empirical argument. Draw implications from reasoning steps within one aspect of the problem that involves empirical entities.	'In January, it's a little warmer. The highest temperature is 10 and the lowest is 1.' 'Monday difference = 6, Tuesday difference = 1, Wednesday difference = 0, Thursday difference = 3, Friday difference = 1, Saturday difference = 4, Sunday difference = 2. Warmest in January.'	2
II	Develop/critique an assumption/claim/statement supported without a sufficient argument. Draw implications with only a simple rationale or external conviction arguments.	'December is snowy, so it is colder, and in January there is no snow.'	1
I	Only an assumption/claim/statement without argument. Draw implications without any argumentation.	'December was the coldest week.' (no arguments – only an assumption) 'January is warmer.'	1
0	No claim or assumption in connection to the question – off track.	Blank	0

Phase 2: Developing the test. The coding guides were developed in an iterative process within the research group (the authors of this paper together with an associate professor) and teachers. Trainee teachers participated as raters in this process, and, to be sure of intercoder reliability, they double-coded 20% of all items in both the pilot study and the three trials in the KiDM project, to fine-tune the categories, all of which needed to be exhaustive. There was 82% consistency between the double-coded items in the KiDM project. Agreement between the scores awarded for TIMSS in 2011 and 2015 ranged, on average, from 92% in science to 98% in mathematics among the year 4 students (Foy et al., 2016), but it is important to note that most TIMSS items are dichotomic, while the current test has polytomous items with up to four different codes, and so the intercoder reliability will inevitably be lower.

The number of test items was relatively small because it was not possible to include many different open-ended items due to it taking a long time for students in years 4 and 5 to answer such questions, for which they had to write their arguments. For ethical reasons, no more items were developed, as the assignment would then be too extensive for these 9-to-11-year-old children to complete. The person separation index would therefore be low in the test and could have been increased by expanding the number of items.

The primary function of the measurement model is to bridge from the empirically found scores to the theoretically constructed model (Wilson & Gochyyev, 2013). The measurement model chosen in this test is the Rasch model. Georg Rasch (1960) introduced a probabilistic model for item response, which gives the probability p of scoring a given category *k* for item *i*, with difficulty δ_{ik} and the ability of the respondent θ; *m* is the maximum score for the item. Here is the partial credit model:

$$p(x_{ni} = k) = \frac{e^{k(\theta_n - \delta_{ik})}}{1 + \sum_{k=1}^{m} e^{k(\theta_n - \delta_{ik})}}$$

Rasch analysis allows us, for example, to compare the difficulty of mathematical reasoning problems while also locating the degree to which individual students have mastered the necessary skill set. This location of reasoning competence and students on the same unidimensional scale allows a fine-grained analysis of which aspects of the reasoning process being analyzed makes one problem more difficult than another. The analysis was done using the RUMM2030 software (Andrich, 2009).

In Figure 2, we see how the probabilities of the different scores (0, 1, 2, 3) develop as a function of the students' ability on item 9 (from Figure 1). When we use a Rasch model as a measuring model, it is important to acknowledge that the answer at a higher location must not only differ from the answer at a lower location, but also be better in a significant way – it must be regarded as superior to the answer at a lower level because, in the Rasch model, we regard students who give an answer at a higher level as showing more ability than those who do not (Bond & Fox, 2015).

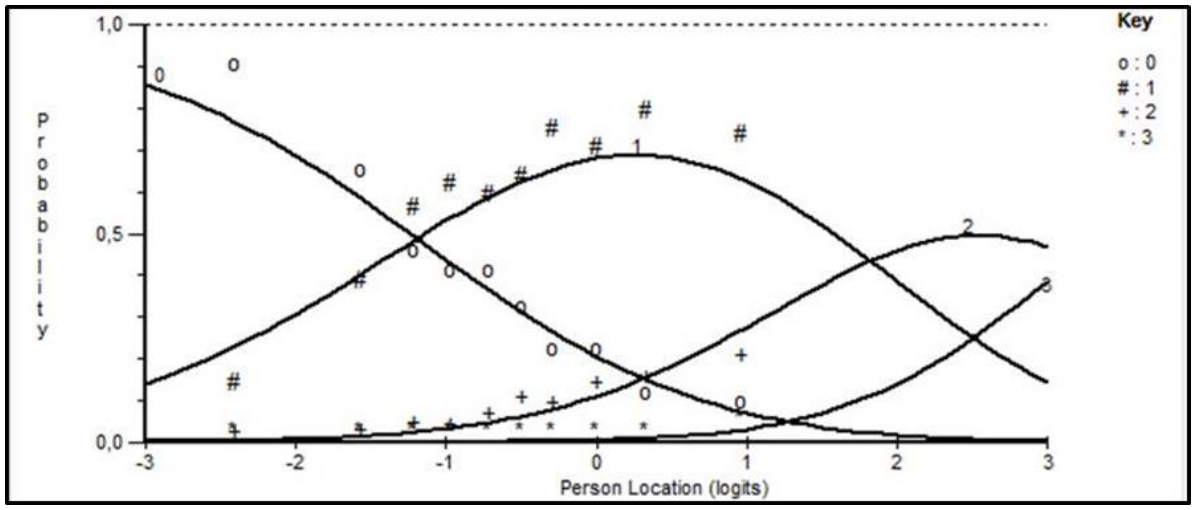


Figure 2. The Probability Curves for Item 9

Based on the students' responses from the pilot test, each item was analyzed using the Rasch model to calibrate the measurement tool, with the aim of eliminating obvious misfits from the test; six items were thus removed after the pilot test.

Phase 3: Content validity after the pilot test. To make sure the items were measuring the specific competence (content validity), two things were done. Together with a mathematical researcher, the first author of this paper worked through all the answers from the pilot schools to determine the extent to which the theoretical purposes were revealed in the students' responses. This means that all the different types of student response were evaluated, and it was determined whether these responses were indicative of the student showing reasoning competence; whether the answers were more about being able to, for example, multiply; or whether they were non-rational guesses that did not show mathematical reasoning.

Furthermore, think-alouds were conducted with two students in which the students' responses were audio recorded and afterwards analyzed to ensure both that the items met the requirement of being understandable to the students' age groups and that the students' responses were desirable answers within the intended content areas. This was to verify that the students, in the process of answering the items, did mathematical reasoning – for example, whether they considered different arguments for their solutions or whether their answers were developed by a using simple skill algorithm or rote learning. As a result, many formulations of the items were changed, and more items were added in the different mathematical areas.

A good measurement process in education will provide for the estimation of just one ability at a time and will not intentionally or unintentionally confuse two or more human attributes into one measure. This is, however, difficult in close-to-reality items. For example, one item might rely on language comprehension to understand a context, while another might focus too much on students' ability to interpret a complex geometrical drawing or algebraic reduction. The resulting score would then be uninterpretable, to the extent that these other abilities are manifested in the students' responses. The important thing for us was that these aspects not overwhelm the role of the desired attributes. In the competence test, it was therefore decided that each of the items should contribute in a meaningful way to the construction of a mathematical competence and that each mathematical competence should have content from all areas – algebra, geometry, and statistics. Some competences had several items from each area.

Results

Construct Validity in the KiDM Test

To validate the construct of the KiDM test, we first conducted four important tests: a monotonicity test, a test for unidimensionality, a test for response dependence, and a measurement of differential item functioning (DIF). The analyses were performed using the test data collected at the baseline (in all three trials). To investigate whether items met the requirement of monotonicity, two different conditions were examined. Monotonicity concerns whether the likelihood of a student responding correctly to an item increases the better the student is. First, we graphically inspected the empirical curve of each item, which shows the proportion of correct responses for student groups, broken down by their skill level and matched with the theoretically expected difficulty of the task. We also inspected the fit-residual statistics (as well as the chi-squared test), which tells how large the deviations are between an item on the empirical and theoretical curves. From a trade-off between the graphical inspection and the fit residual, four were excluded from the scale because the deviations between the empirical and theoretical curves were too large.

We also investigated whether the items fit to the dimension (a unidimensionality test) and did not overdiscriminate or underdiscriminate. With residuals smaller than −2.5, we found 11 items that overdiscriminated, while there were 22 items that underdiscriminated, with residuals larger than 2.5. For students with low ability, an item with a large discrimination would be the most difficult, but the same item would be the easiest for a person with high ability. Other items would overtake the difficulty of the high-discriminating item for the high-ability students, and the item difficulty would depend on the sample of students. We therefore deleted the 10 most extreme items.

To investigate whether there was response independence between the individual items, we looked at the residual correlations between each item pair. A residual correlation indicates whether the answer to a task affects the probability of responding correctly to another task. One item pair had a residual correlation above 0.3 and was eliminated from the test; this was a sign of a breach of the requirement for local independence between the items. The eliminated item was selected from a closer examination of which of the two items in the item pair had the highest residual correlation with other items.

Finally, to ensure that particular groups were not prejudiced, the DIF was tested. This involves controlling for respondents' overall ability and identifying whether an item favors one group of respondents over another. Some groups might perform differently overall, but it is not desirable that students from one group find it harder or easier to answer an item correctly than students of equal ability from another group. We checked for the DIF according to grade level and whether the students attended an intervention school or a control school at the time the student took the test.

As assumed, we found no significant differences caused by the students' belonging to the intervention or control group, because the schools were randomly selected, although we found some unproblematic differences in the grade level. Figure 3 shows that year 5 students (horizontal lines) had some advantages over year 4 students (sloping lines) because they were taught mathematics during their additional year of schooling. This did not, however, pose any DIF issues. We also saw some similarly unproblematic differences between the three trials: Trial 2 of the competence test had the highest mean (−.571), while the mean for Trial 1 was −.744 and the mean for Trial 3 was −.915. This is unsurprising, as Trial 2 took place in the spring, so the classes in that trial had attended school half a year longer than the classes in Trials 1 and 3.

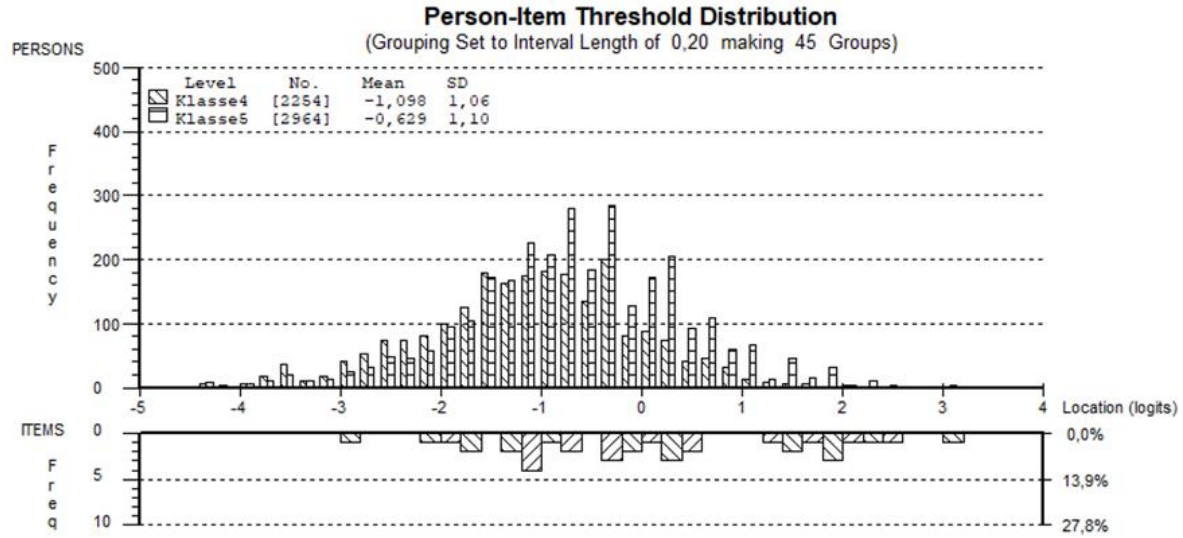


Figure 3. Person–item Threshold in the Competence Test in School Years 4 and 5

Validating through Item Maps

By using the Rasch model to analyze the test answers, it was also possible to use an item difficulty map (Van Wyke & Andrich, 2006) (see Figure 4). This map shows the relationship between items and students, taking advantage of the fact that persons and items are measured on the same scale. The logit scale, which is the measurement unit common to both person ability and item difficulty, is displayed down the middle of the map. Because the logit scale is an interval-level measurement scale, equal distances at any point on the vertical scale represent equal amounts of cognitive development.

On the item map, the distance of the step from the bottom of the path represents its difficulty – items closer to the bottom are easier, those further up are more difficult. To the left, the students are indicated by small crosses. In Figure 4, one cross indicates 27 students. Students closer to the bottom have less ability than students at the top. SU is an abbreviation for the Danish word *Scorede Undersøgende-opgave*, which can be translated into *scored inquiry-based items.*

An item map can help researchers identify the strengths and weaknesses of an instrument, such as if some test items are measuring the same part of the variable or if there are areas of the tested variable that are missing from the test due to a lack of items with different levels of difficulty. In developing this test, our aim was to place enough stepping-stones along the path between little development and much development to represent all the points useful for our testing purposes; to do that, we needed to collect observations from enough suitable persons, but we also needed to have enough items.

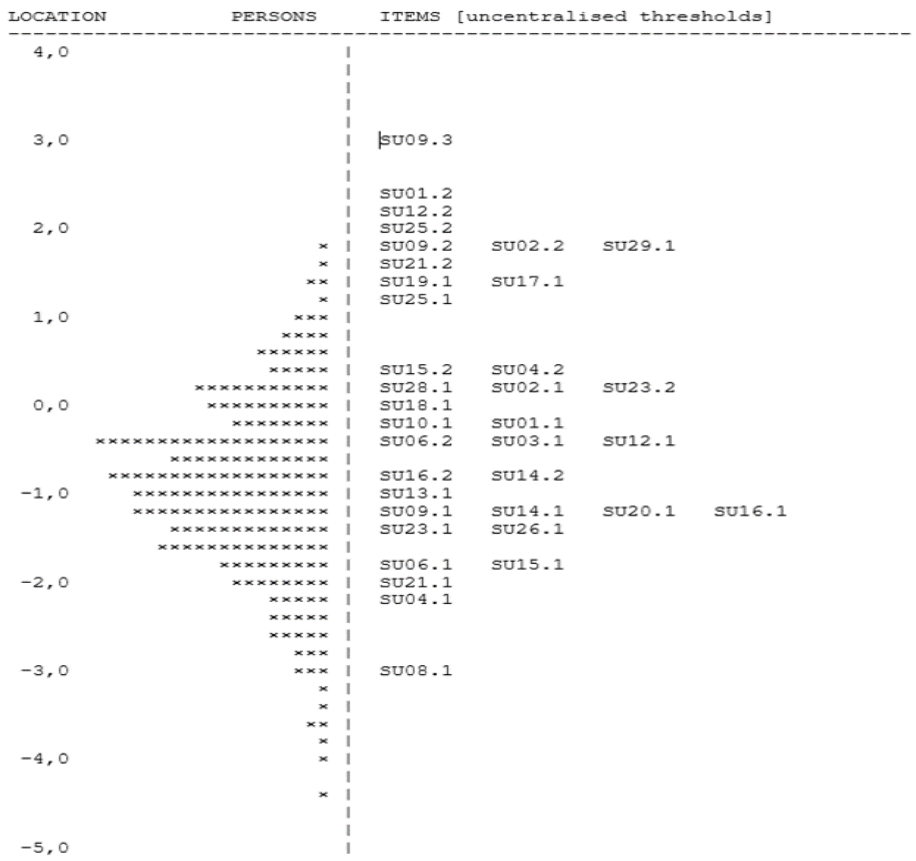


Figure 4. Item Map for the Items in the Competence Test

In the item map in Figure 4, we see that there are some gaps in the items along the vertical line, which may indicate that we lack some items with a specific level of difficulty. We can also use this item map to study the validity of the test by exploring whether there is consistency between the theoretically developed levels and the empirically developed item map. The idea of the item map is that each student will progress along the steps as far as their ability allows – a student will master the steps until the steps become too difficult. How far any student moves along the pathway will be our estimate of the student's ability development. We must, however, remember that the Rasch model is probabilistic and therefore does not mean the students will correctly answer all the items up to the point at which they are placed and incorrectly answer all the items above it. Rather, the item map suggests that a person has a more than 50% probability of responding correctly to tasks below their skill level and a less than 50% probability for tasks above their skill level. The expected outcome will therefore be different from the actual outcome.

Such comparisons facilitate an assessment of construct validity by providing evidence that the instrument is measuring in a way that matches what theory would predict. In Figure 5, the item map from Figure 4 has been coded, which is shown with different colors. The items that do not have a specific focus on reasoning competence have been removed. The black boxes with white numbers are items for which the students made correct assumptions and claims but without any argumentation (Locations I and II in Table 2), and the grey boxes with black numbers are items for which the students made argumentations for their assumptions or for critiques (Locations III, IV, and V in Table 2). Figure 5 clearly shows that it is more difficult to produce arguments than to simply present assumptions or elicit claims in the test items.

The item at the top, SU9.3, is shown to be the most difficult item. In this item, code 3 (from Table 2) is where the students had the chance to show they could draw implications by linking pieces of information from separate aspects of the problem to make arguments (Location V). In the beginning of the test-development phase (phase 2), more items in the test were possible to be answered in this way, but there simply were not enough students answering correctly at this difficulty level for this coding to be retained in the model, so we had to remove these codes from a few items and rethink the code boundaries or combine some codes. This might simply indicate that this is a very difficult way to reason for students in school years 4 and 5.

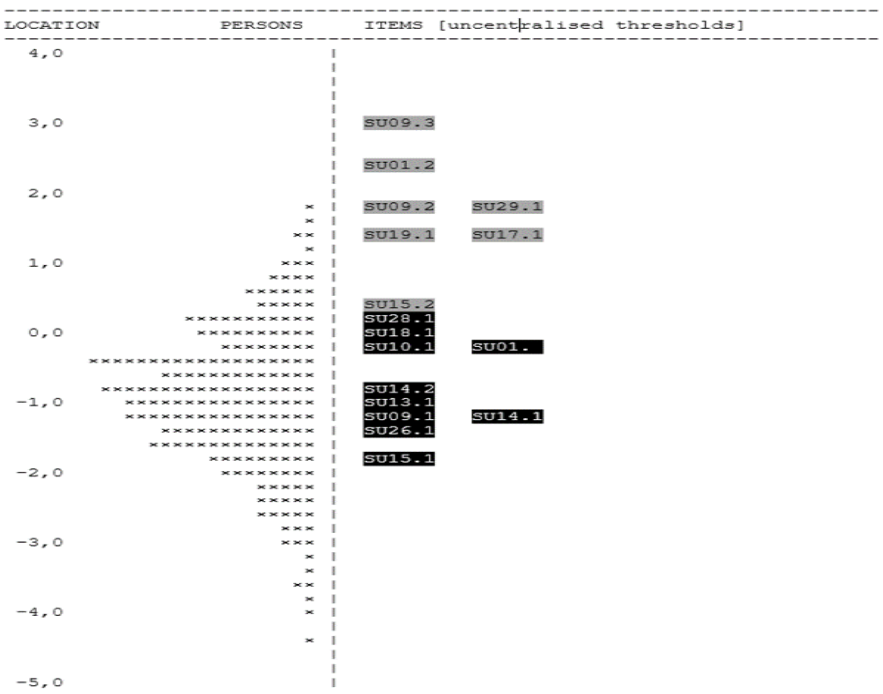


Figure 5. Item Map with Codes Indicating Argumentation

In Figure 6, a new coding has been made: the grey boxes with black numbers are complex items and the black boxes are non-complex items. A complex item is defined here as one in which the students need to calculate more than one result before being able to make an argument or claim. In the Figure, we see that most of the complex items are at the top.

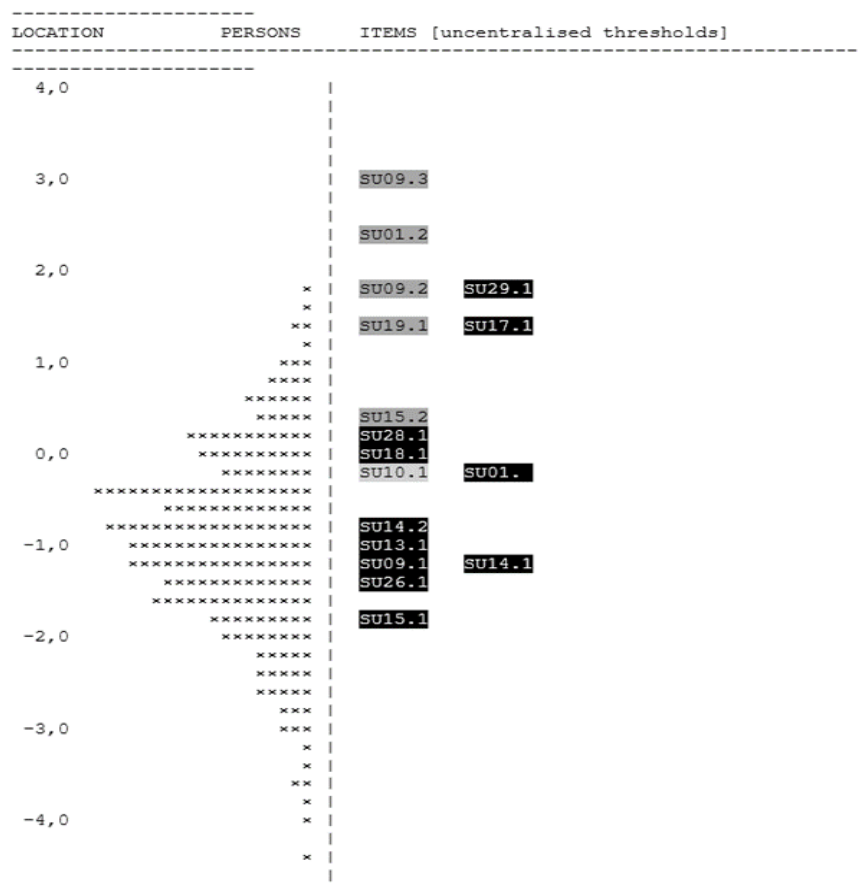


Figure 6. Item Map with Codes Indicating Complexity of Items

Discussion of Validation

During the analysis, we coded for other aspects in the item map as well. For example, we coded items in which the students developed their own claims compared to items in which the students only had to understand existing claims. This coding, however, only showed that one was not significantly harder than the other, but we observed that some items can probably be solved using different methods and that this may call for the activation of the competences in different combinations or at different levels, which may have changed the item map in another direction. It is therefore recognized that some degree of dissimilarity in the outcomes is inevitable.

These kinds of comparison help to facilitate an assessment of validity by providing evidence that the instrument is measuring in a way that matches what the theory would predict. To test students' ability in mathematical reasoning and to be aware of how it grows and becomes more sophisticated over time is interesting for multiple reasons. It is interesting for the students to be aware of where they are heading, but it is also important for researchers in the development of curricula, teaching materials, and tests and other evaluation tools.

Testing Mathematical Competences in the KiDM Project

The final number of items after calibrating the test was 23, of which 13 were focused on reasoning. This number may be considered large when one contemplates that open-ended items require students to write a short passage, but in the context of the Rasch model's requirements, the number is relatively small. After collecting baseline and endline data in the control and intervention schools in all three trials, the final results of the KiDM test indicate that it is not possible to measure any positive effect in student mathematical competences in the intervention schools compared to the control schools, with a statistical significance level of 0.05. This means that we cannot rule out that the differences between the intervention and control group levels of mathematical competence were coincidental.

There may be several reasons for this. The intervention, over a relatively short period, may not trigger any significant changes for the students at the intervention schools compared to the control schools, who may also have been taught competences in mathematics. Another issue is that, even though there were many students included in the trial, the randomization was carried out at school level – cluster-corrected standard errors were used, taking into account that students who attended the same school were not independently drawn, but the students were clustered in schools. The number of intervention schools was 38, with 45 control schools, which is a relatively small number and may also have had an effect on the possibilities for finding a significant result.
However, we can identify two reasons why the test may not have worked as desired. Firstly, the fitting of the items to the construct meant that a relatively high percentage of the items were removed, resulting in relatively few items (23) that fitted the model; this relatively small number of items would affect the results. Secondly, many of the items were located at the difficult end of the scale, compared to the students, which means that many students did not respond correctly to any of the items.

Conclusion and Implications

We have shown that there is some consistency between our empirically developed item map and the theoretically developed map, but we also found some limitations. One disadvantage of the data is that, in this test, the items are separate components – it is unclear whether they can ever add up to higher-order thinking in which knowledge and skills can be detached from their contexts or their practice and use. Is it at all possible to measure a skill component in one setting that is learned or applied in another? This means that, perhaps, we cannot validly assess a competence in a context very different from the context in which it is practiced or used (Resnick & Resnick, 1992); the test might have been improved if this had been even more in focus during development.

Moreover, we also see some limitations in connection with students who have difficulty writing. Although the tasks could be read out by a computer, the year 4 students especially may have had trouble arguing their answers in writing. Here, it could have been interesting for the students to be able to draw their answers, which, unfortunately, was not possible in our setting. Finally, a measure of concurrent validity is missing, and it could have been interesting to compare the students' scores in this test to their scores in the national test, to see if the scores were predictive; however, this was also not possible in our setting.

The item maps are very interesting from a mathematical educational research perspective, because they can inform the discussion about what students find easier or more difficult using empirical findings and not just expert opinion. Many different codings could be made; for example, are items with context or without context more difficult? Do illustrative pictures make an item easier? However, to do studies like this, we need to be very aware of how the items are developed, to ensure that these aspects can be measured.

In conclusion, this paper presents a serious attempt to answer how an achievement test to measure reasoning competence in mathematics can be designed, developed, and tested, and the KiDM test is a serious suggestion for a more open-ended test that measures mathematical competence. Our purpose in this paper is to encourage mathematical education researchers to not only criticize achievement tests, but also to focus on how we can develop new alternatives to measuring other competences or to improve already developed tests, given that testing seems to be here to stay.

References

Andrich, D., Sheridan, B., & Luo, G. (2009). RUMM2030: Rasch unidimensional models for measurement. Perth, Western Australia: RUMM Laboratory.

Ball, D. L., & Bass, H. (2003). Making mathematics reasonable in school. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 27–44): Reston, VA: National Council of Teachers of Mathematics.

Biggs, J. (2011). *Teaching for quality learning at university: What the student does*. London, UK: McGraw-Hill Education.

Black, P. J. (1998). *Testing, friend or foe? The theory and practice of testing and assessment*. London, UK: Falmer Press.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

Brousseau, G., & Gibel, P. (2005). Didactical handling of students' reasoning processes in problem solving situations. In C. Laborde, M.-J. Perrin-Glorian, & A. Sierpinska (Eds.), *Beyond the apparent banality of the mathematics classroom* (pp. 13–58). Boston, MA: Springer US.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302.

Duval, R. (2007). Cognitive functioning and the understanding of mathematical processes of proof. In P. Boero (Ed.), *Theorems in school* (pp. 137–161). Rotterdam, Netherlands: Sense.

EMS. (2011). Do theorems admit exceptions? Solid findings in mathematics education on empirical proof schemes. *Newsletter of the European Mathematical Society, 81*, 50–53.

Foy, P., Martin, M. O., Mullis, I. V. S., Yin, L., Centurino, V. A. S., & Reynolds, K. A. (2016). Reviewing the TIMSS 2015 achievement item statistics. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 11.11–11.43). Retrieved from http://timss.bc.edu/publications/timss/2015-methods/chapter-11.html

Goetz, T., Preckel, F., Pekrun, R., & Hall, N. C. (2007). Emotional experiences during test taking: Does cognitive ability make a difference? *Learning and Individual Differences, 17*(1), 3–16.

Hanna, G., & Jahnke, H. N. (1996). Proof and proving. In K. C. A. Bishop, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (pp. 877–908). Dordrecht, Netherlands: Kluwer Academic Publishers.

Harel, G., & Sowder, L. (1998). Students' proof schemes: Results from exploratory studies. In A. H. Schoenfeld, J. Kaput, & E. Dubinsky (Eds.), *Research in collegiate mathematics education III* (pp. 234–283): Providence, RI: AMS.

Johnson, R. B., & Christensen, L. B. (2014). *Educational research: Quantitative, qualitative, and mixed approaches* (5th ed.). Thousand Oaks, CA: Sage.

Knuth, E. J. (2002). Secondary school mathematics teachers' conceptions of proof. *Journal for Research in Mathematics Education, 33*(5), 379–405.

Larsen, D. M. (2017). *Testing inquiry-based mathematic competencies. Short communication*. Paper presented at the Merga Conference 40, University of Monash, Melbourne, Australia.

Lithner, J. (2008). A research framework for creative and imitative reasoning. *Educational Studies in Mathematics, 67*(3), 255–276.

Logan, T., & Lowrie, T. (2013). Visual processing on graphics task: The case of a street map. *Australian Primary Mathematics Classroom, 18*(4), 8–13.

Niss, M., Bruder, R., Planas, N., Turner, R., & Villa-Ochoa, J. A. (2016). Survey team on: Conceptualisation of the role of competencies, knowing and knowledge in mathematics education research. *ZDM, 48*(5), 611–632.

Niss, M., & Jensen, T. H. (2002). *Kompetencer og matematiklæring: Idéer og inspiration til udvikling af matematikundervisning i Danmark* [Competencies and mathematic learning: Ideas and inspiration to development of teaching in mathematics in Denmark] (Vol. 18). Copenhagen, Denmark: Danish Ministry of Education.

Nunes, T., Bryant, P., Evans, D., & Barros, R. (2015). Assessing quantitative reasoning in young children. *Mathematical Thinking and Learning: An International Journal*, *17*(2–3), 178–196.

Nunes, T., Bryant, P., Evans, D., Bell, D., Gardner, S., Gardner, A., & Carraher, J. (2007). The contribution of logical reasoning to the learning of mathematics in primary school. *British Journal of Developmental Psychology*, *25*(1), 147–166.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedogogiske Institut (Chicago: University of Chicago Press, 1980).

Reid, D. A., & Knipping, C. (2010). *Proof in mathematics education, research, learning and teaching*. Rotterdam, Netherlands: Sense Publishers.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37–75). Dordrecht, Netherlands: Springer.

Shaughnessy, M., Barrett, G., Billstein, R., Kranendonk, H., & Peck, R. (2004). *Navigating through probability in grades 9-12*. Reston, VA: National Council of Teachers of Mathematics.

Stylianides, A. J. (2007). The notion of proof in the context of elementary school mathematics. *Educational Studies in Mathematics*, *65*(1), 1–20.

Stylianides, A. J., & Harel, G. (2018). *Advances in mathematics education research on proof and proving: An international perspective*. New York, NY: Springer.

Stylianides, G. J. (2008). An analytic framework of reasoning-and-proving. *For the Learning of Mathematics, 28*(1), 9–16.

Stylianides, G. J., & Stylianides, A. J. (2017). Research-based interventions in the area of proof: The past, the present, and the future. *Educational Studies in Mathematics [Special Issue]*, *96*(2), 119–274.

Van Wyke, J., & Andrich, D. (2006). A typology of polytomously scored mathematics items disclosed by the Rasch model: Implications for constructing a continuum of achievement. *Unpublished report, Perth, Australia*.

Wilson, M., & Gochyyev, P. (2013). Psychometrics. In T. Teo (Ed.), *Handbook of quantitative methods for educational research* (pp. 3–30). Rotterdam, Netherlands: Sense.

Yackel, E., & Hanna, G. (2003). Reasoning and proof. In W. G. Martin & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 227–236). Reston, VA: National Council of Teachers of Mathematics.