An assessment of non-standardized tests of mathematical competence for Norwegian secondary school using Rasch analysis

Morten Klegseth, Eivind Kaspersen, Trygve Solstad

An assessment of non-standardized tests of mathematical competence for Norwegian secondary school using Rasch analysis

Morten Riise Klegseth, Eivind Kaspersen and Trygve Solstad

Norwegian University of Science and Technology, Faculty of Social and Educational Sciences, Department of Teacher Education, Trondheim, Norway
Morten-Riise.Klegseth@ou.trondheim.kommune.no; eivind.kaspersen@ntnu.no; trygve.solstad@ntnu.no

Do non-standardized, publisher-provided tests for lower secondary school provide valid and reliable measures of mathematical competence? We analysed a sample of items pooled from tests accompanying three different Norwegian textbooks using Rasch analysis. The pooled sample of items was found to be sufficiently unidimensional for measuring function competence, with four strands of sub-competencies in accordance with theory. The competence associated with an increasing difficulty of items could be qualitatively characterised by shifts from a) identifying through constructing to reasoning about representations, b) using visual to using algebraic representations, and c) local to global interpretations of functions. While the individual tests differed substantially in the distribution of items across strands of mathematical competence, minor adjustments to the combined instrument were sufficient for providing a valid and reliable measure of mathematical competence.

Keywords: Competence, assessment, rasch, functions.

Introduction

Background

Teachers rely strongly on written tests when determining final grades for secondary school pupils (Brookhard, 1994). In Norway, final grades are to a large degree based on results on non-standardized midterm exams and shorter tests provided by textbook publishers (Prøitz & Borgen, 2010). Publisher-provided tests are composed by experienced textbook authors without explicit reference to a theoretical framework for the mathematical competence the tests aim to assess. Because mathematical functions are central to the field of mathematics and to the Norwegian secondary school curriculum, and because they are typically introduced in the 10th grade, the topic constitutes a relevant and convenient source of information about how such tests are composed and what they measure. In this study we asked a) how competence in linear functions is operationalised in tests accompanying Norwegian mathematics textbooks, and b) to what extent these tests provide valid and reliable measures of mathematical competence.

Theoretical framework for competence in mathematical functions

As a starting point for describing mathematical competence, we took the widely used Danish KOM model, which distinguishes eight partially overlapping mathematical *competencies* (Niss & Jensen, 2002). Briefly, these competencies are i) Mathematical thinking, ii) Mathematical problem solving and -posing, iii) Mathematical modelling, iv) Mathematical reasoning, v) Handling mathematical representations, vi) Handling mathematical symbols and formalisms, vii) Mathematical communication, and viii) Using aids and tools.

The tests we analysed were dominated by questions concerning transformations of different representations of functions, warranting a further characterization of this competency. Representations form the foundation of many theoretical frameworks for mathematical competence and have been central to describing competence in functions. O'Callaghan (1998) presents a model with four main components: 1) *Modelling*, the transformation from a problem situation to a mathematical representation using functions, 2) *Interpretation*, the transformation from a mathematical representation of a function to the description of a problem situation, 3) *Translation* between representations of functions, like symbols, tables, and graphs, 4) *Reifying*, the creation of a mental object from what was initially perceived as a process or procedure, and 5) *Procedural skills* for operating within a representation system.

Levels of competence

The present manuscript focuses on the role that textbook tests have in determining students' final grades. These tests are typically administered after each mathematical topic has been covered in class and can be considered high stakes in the sense that they collectively make up part of the basis for a teacher's end-of-school assessment. However, these tests often serve formative as well as summative purposes. While the summative aspect differentiates students according to their levels of competence, qualitative characterizations of each level of competence within a competency can both address issues of test validity and be useful in a formative perspective on assessment.

The perceived difficulty of a question about mathematical functions has been shown to depend on the cognitive demand of providing a valid answer to a problem. In particular, *interpreting* or *recognizing* properties of a given representation or statement is easier than *recalling* or *constructing* a solution to a problem when a target representation is not given. *Explaining* why a solution is valid typically requires the student to explicate relations between multiple representations and is perceived as more difficult than identifying and constructing valid representations (Leinhardt et al., 1990; Nitsch et al., 2015).

Representations of functions and transitions between representations can be interpreted from a local or global perspective. Whereas local interpretations of a function involve accessing single values of the representation, global interpretations involve reasoning about how the function behaves as a whole or within certain intervals of the domain. Global interpretations are important for accessing more advanced mathematics and are associated with higher levels of mathematical competence (Leinhardt et al., 1990; Gagatsis & Shiakalli, 2004; Duval, 2006; Bossé et al., 2011).

These perspectives on what characterizes different levels of competence served to aid our analysis of the test items from the non-standardized textbook tests.

Methods

Selection of test items

Three tests accompanying 10th grade textbooks from three major Norwegian publishers were selected as a source of common test items in Norwegian schools. When two or more tests contained very similar items, only one item was selected for our instrument. Items that did not address the subject of linear functions were excluded from the study. One item was excluded because it could not be faithfully translated into digital form. Items requiring a global interpretation of functions were missing from the original tests, and we added two items in order to assess this category of competence. After this selection process our pooled test consisted of 31 test items.

Modification of test items

After a pilot study, Rasch analysis identified some items as unreliable. In particular, multiple choice items did not provide good fits to the Rasch model and were converted into explanation items. One original item used specific numbers that produced ambiguous answers, and a new set of more suitable numbers was chosen. For a few items, we adjusted the specific numbers used in order to obtain a more uniform distribution of item difficulties in the instrument as a whole.

Categorization of test items

The test items were categorized according to the theoretical frameworks discussed in the introduction. While no items fell into the ‘reification’ category, several items asked about specific concepts. Reification was therefore substituted with a separate category for Concepts, and we used the following categories for the analysis: Interpretation, Translation, Modelling, Concepts, Coordinates and Others. The three first categories were generated directly from the theory of competence for functions. Concepts can be considered part of mathematical thinking in the KOM framework (Niss & Jensen, 2002, p. 47), and coordinates can be considered a part of “symbols and formalisms” in the KOM framework (Niss & Jensen, 2002, p. 58). Items in the “others” category were excluded from the study as they were not directly related to competence in linear functions, like items requiring competence in nonlinear functions, solving equations, and general competence with digital tools.

An anticipation of item difficulty was estimated (“easy”, “medium”, or “difficult”) for each item based on whether the item required a) identification or interpretation of a given solution, b) construction of a valid solution, or c) an explanation for a mathematical statement (Nitsch et al., 2015; Leinhardt et al.,1990). The anticipated difficulty was adjusted according to whether the item required a) a local or global interpretation of the given function, and b) one or multiple transformations between representations of the given function.

Participants

A convenience sample of fourteen school classes with a total of 250 tenth grade pupils from 5 out of 13 secondary schools in the city of Trondheim, Norway, participated in the study. All pupils had completed classroom instruction in linear functions between one week and two months before they participated in the study. Participation was voluntary, and all answers were anonymous.

Data collection

The test items were digitized and answers to the items were collected using a web platform developed at the Department of Teacher Education, NTNU. After a 5-minute presentation of the testing tool and informed consent, pupils had 55 minutes to complete the test. Test items were presented in randomized order. If no answer was given to a test item, the answer was coded as "missing data".

Analysis

As most items in the tests asked for a single correct answer, each item was scored either 0 or 1 point (dichotomous model). Quantitative data was analysed with the Rasch model (Rasch, 1960) using the Winsteps software (Linacre, 2017). In the Rasch model, the probability that person *v* scores 1 point on item $\boldsymbol{\iota}$ depends on the difference between the ability of person *v*, $\boldsymbol{\beta_v}$, and the difficulty of item $\boldsymbol{\iota}$, $\boldsymbol{\delta_\iota}$, according to

$$\boldsymbol{P\,\{X_{v\iota} = 1|\beta_v, \delta_\iota\} = \frac{e^{(\beta v\,-\,\delta\iota)}}{1 + e^{(\beta v\,-\,\delta\iota)}}}$$

Winsteps implements the joint maximum likelihood estimation (JMLE) algorithm for estimating the parameters of this model, and principal component analysis (PCA) of normalized residuals for investigating the dimensionality of the dataset.

Validity

Assessment of the validity of the instrument was based on the framework presented in Wolfe and Smith (2007) which expands on Messick (1995). Here, we considered the following six aspects of validity: *i) Content* (e.g. relevance, representativeness, and technical quality), *ii) substantive* (e.g. theoretical foundation), *iii) structural* (e.g. evidence of unidimensionality), *iv) generalizability* (e.g. generalization across sample and context), *v) consequential* (e.g. fairness and possible biases), and *vi) interpretability* (e.g. the relationship between quantitative measures and qualitative meaning).

Results

Classification of item competence

We collected test items from three publisher-provided tests in functions for Norwegian secondary school. The items were classified into five competence categories and assigned an anticipated level of difficulty (depending on cognitive demand, number of representational transformations, and local vs global perspective on functions). While the combined set of items covered a broad range of competence categories, individual tests differed substantially in their emphasis on each competence category (Table 1). In particular, tests B and C put opposite emphasis on interpretation, translation and concepts, while items from test A were more evenly distributed among the categories.

Test	Concept	Interpretation	Translation	Modelling	Coordinates	Excluded
A	4	4	5	2	1	5
B	0	7	1	3	1	3
C	4	1	6	4	4	5

Table 1: Number of items in each competence category for each of the tests in the study.

Measurement properties

In general, the data fit well with the Rasch model, contributing to the instrument's content validity (Wolfe & Smith, 2007). Person reliability, analogous to Cronbach's alpha, was 0.86. Item infit mnsq was 1.01 ± 0.18 (mean ± std) and item outfit mnsq was 0.9 ± 0.34 (mean ± std), which means that the variance in pupils' responses to single items generally fit well with the Rasch model (infit mnsq = outfit mnsq = 1).

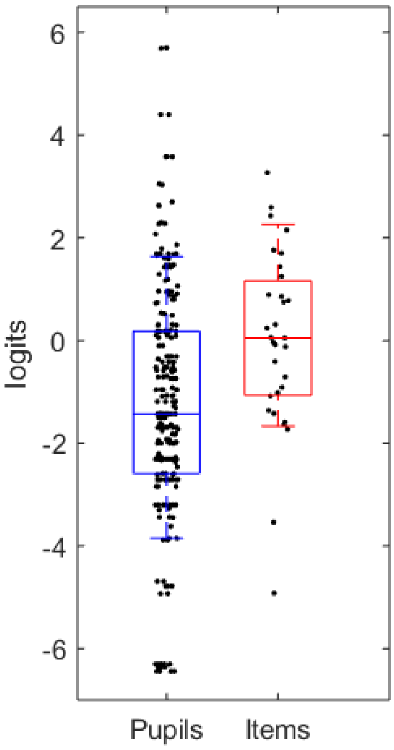

Figure 1. Competence level of pupils (blue boxplot) and item difficulty (red boxplot) on the same logit scale. Boxplots indicate 10th, 25th, 50th, 75th and 90th percentiles.

The item difficulties ranged from −4.9 to 3.3 logits, but the distribution of items was uneven with 28 out of 31 items between −1.7 and 2.6 logits (Figure 1, red boxplot). This contrasted with the distribution of pupil achievement level: only about half (52%) of the pupils achieved within this range of item difficulties, and almost 75% of pupils achieved below the mean item difficulty (Figure 1, blue boxplot). The low attainment might be related to the time of testing and the perceived distance to final exams.

Only two items were positioned to discriminate achievement levels below the −1.7 logit mark where 42% of pupils were measured. These items might be considered at the entry level to competence with functions, and as a tool for summative assessment the instrument distinguishes well between pupils above that competence level. However, the large gaps between easy items, producing an abundance of low test scores, leaves a large proportion of students without feedback about their competence level beyond an unintended subtext of failure. From the view of formative assessment, the scarcity of easy items detracts from the instrument's content and consequential validity (Wolfe & Smith, 2007).

Item difficulty was largely invariant to the pupils' achievement level, as determined by comparing the difficulty level of each item between the highest achieving and lowest achieving pupils. Pupil achievement level affected the difficulty of only 2 out of the 31 items (at the p = 0.0016 level; Bonferroni corrected for multiple comparisons from p = 0.05), contributing the instrument's generalizability validity (Wolfe & Smith, 2007). The first of these items favoured high-achieving pupils and was the only item involving a function with negative slope. The second item favoured low-achieving students by unintentionally allowing the zooming in on a graph to read off the solution directly rather than reasoning about it.

Empirical categories of competence

The competence categories were taken from theoretical frameworks for competence with functions. To investigate if these categories could be identified in the empirical data, we conducted a PCA on standardized residuals of the data (Linacre, 2017). PCA identified two contrasts with potential subdimensions. The first contrast (eigenvalue = 2.4) clearly separated Interpretation items (the six items with the highest positive loading) from Translation items (the six items with the highest negative loading). In addition, all 13 items in the cluster with negative loading included symbolic expressions. The second contrast (eigenvalue = 1.9) separated the full set of five Concept items together with the single coordinate item from the main dimension of the instrument. The Modelling items did not deviate from the main dimension defining competence levels for linear functions.

While 49% of the variance in the data could be explained by the measures, the additional variance explained by the four clusters combined was around 7%, which can be usefully considered sub-dimensions of the main variable. Taken together, the instrument can be considered unidimensional for measuring competence in linear functions, which adds to its structural validity. At the same time, the dimensionality analysis lends empirical support to the notion that competence in linear functions is composed of four strands, each dominated by one of the four main competence categories Concept, Interpretation, Translation, and Modelling. This correspondence between empirical clusters and theoretical foundation speaks to the instrument's substantive validity.

Empirical levels of competence

What is the qualitative meaning of the quantitative measure along the competence scale?

First, the distribution of item difficulty did not differ significantly between competence categories, either in variance (p = 0.15; Levene's test) or mean difficulty (p = 0.13; one-way ANOVA; Figure 2). However, the two most difficult items in the Interpretation category were added to the original items because a global perspective of functions was missing from the original tests. Without these two items, the mean difficulty of Interpretation items was significantly lower than for items in the other categories (p = 0.01; one-way ANOVA).

Two items in the Interpretation category stood out as easier than other items. These items asked pupils to read off a value in a coordinate system and count the number of constant parts of a graph. Arguably, both items could be classified as prerequisite for, rather than part of, graphical representations of functions. At the same time, the competence of as much as 40% of the students were estimated to be within this prerequisite level of function competence, challenging the validity of the test as a formative tool. Beyond the fact that both students and teachers are deprived of the positive effects of feedback for learning, close-to-zero test scores counteract a fair assessment of competence and are potentially harmful to students' motivation for learning (e.g. Schinske & Tanner, 2014). To fulfil its role as a formative tool and guide low-attaining students towards proficiency with functions, the test set should be supplemented with items about prerequisite competencies for functions.

Second, the predicted item difficulty fit well with the empirically estimated item difficulties (different colours in Figure 2). Three exceptions were of interest. These three items were measured to be more difficult than anticipated from theory and shared a pattern in the kinds of mistakes pupils made in answering them: The second and third most difficult Concept items (see Figure 2) asked pupils to identify the slope of a function expression. Most pupils either a) mistook the constant term for the slope, or b) included the variable with the slope in their answers ($\frac{x}{4}$ and $3x$). These very common mistakes resulted in higher-than anticipated measures for these two items. The unexpectedly high estimate of the coordinate system item was also due to a widespread mistake. The item asked pupils to plot a line between two given points, and a surprisingly large proportion of pupils interpreted the two coordinates as four points in the coordinate system instead of two.

Third, 8 out of the 9 items with difficulty less than -1 logit did not concern symbolic expressions. The one item that did was the easiest Modelling item, and asked pupils to form an equation rather than a function from a written context. The item could be solved using an additive strategy without noticing a functional relationship between two variables. -1 logit seems to mark a threshold above which competence with the algebraic symbol system for functions is required. 58% of the pupils in this study scored below the level requiring competence with symbolic expressions for functions.

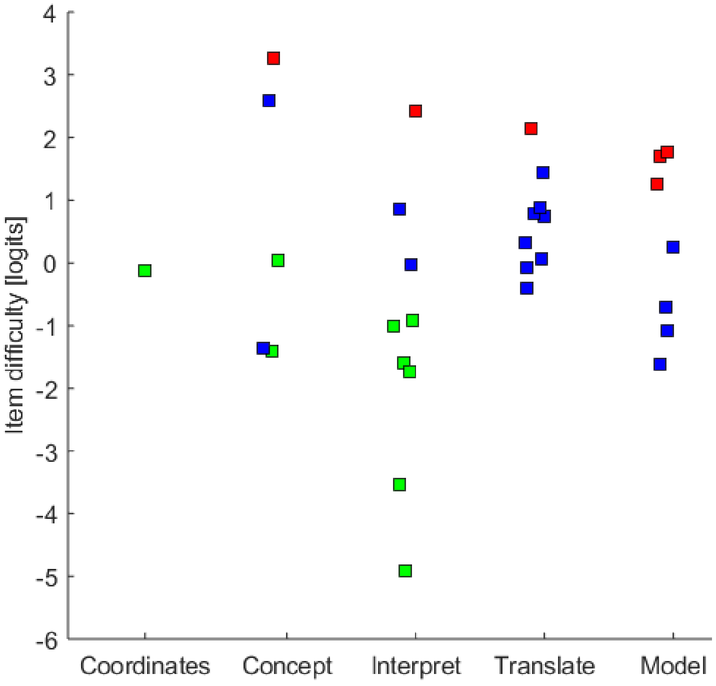

Figure 2: Item difficulty plotted by competence category on x-axis and coloured by anticipated difficulty. Green: easy item, blue: intermediate item, red: difficult item.

The qualitative stratification of items along the difficulty scale gives the instrument interpretability validity (Wolfe & Smith, 2007). The stratification is also useful for formative assessment, but only for pupils that have acquired a minimum level of competence with functions.

Conclusions

We have presented an analysis of a test pooled from three publisher-provided tests of competence in linear functions. The analysis shows that, with minor modifications to some test items, the test set as a whole is a reliable and valid measurement instrument that can be considered one-dimensional for its intended purpose yet consists of empirically identifiable strands of competence that correspond closely to the theoretical framework for mathematical competence outlined in the introduction.

The study suggests that if items are sampled in a balanced manner across both different subdimensions and difficulty levels according to the theoretical framework, calibrated standardized tests might not be necessary to obtain reliable and valid summative assessment of mathematical competence on small scale tests for secondary school. However, adding items on topics prerequisite to competence with functions would strengthen the instrument’s value as a formative tool.

Acknowledgements

These results were previously presented in Norwegian in the Master thesis of Morten R. Klegseth.

References

Bossé, M. J., Adu-Gyamfi, K., & Cheetham, M. R. (2011). Assessing the difficulty of mathematical translations: Synthesizing the literature and novel findings. *International Electronic Journal of Mathematics Education*, *6*(3), 113–133.

Brookhart, S.M. (1994). Teachers’ grading: Practice and theory. *Applied Measurement in Education*, *7*(4), 279–301.

Duval, R. (2006). A cognitive analysis of problems of comprehension in a learning of mathematics. *Educational Studies in Mathematics*, *61*(1), 103–131.

Gagatsis, A., & Shiakalli, M. (2004). Ability to translate from one representation of the concept of function to another and mathematical problem solving. *Educational Psychol.*, *24*(5), 645–657.

Leinhardt, G., Zaslavsky, O., & Stein, M. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. *Review of Educational Research, 60*(1), 1–64.

Linacre, J. M. (2017). *Winsteps® Rasch measurement computer program. User’s Guide.* Beaverton, OR: Winsteps.com.

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.

Niss, M., & Jensen, T. H. (eds.) (2002). Kompetencer og matematiklæring: Idéer og inspiration til udvikling af matematikundervisning i Danmark. *Uddannelsesstyrelsens temahæfteserie 18.* Copenhagen, Denmark: The Danish Ministry of Education.

Nitsch, R., Fredebohm, A., Bruder, R., Kelava, A., Naccarella, D., Leuders, T., & Wirtz, M. (2015). Students' competencies in working with functions in secondary mathematics education–empirical examination of a competence structure model. *International Journal of Science and Mathematics Education*, *13*(3), 657–682.

O'Callaghan, B. R. (1998). Computer-intensive algebra and students' conceptual knowledge of functions. *Journal for Research in Mathematics Education, 29*(1), 21–40.

Prøitz, T.S., Borgen, J.S. (2010*). Rettferdig standpunktvurdering–det (u)muliges kunst?: Læreres setting av standpunktkarakter i fem fag i grunnopplæringen*. Oslo, Norway: NIFU STEP Report, 16/2010.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. *Copenhagen, Denmark: Danish Institute for Educational Research*. Expanded edition, 1980. Chicago, IL: University of Chicago Press.

Schinske J., Tanner K. (2014). Teaching more by grading less (or differently). *CBE—Life Sciences Education*, *13*(2), 159–166.

Wolfe, E. W., & Smith, J. E. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II–validation activities. *Journal of Applied Measurement*, *8*(2), 204–234.