Int J of Sci and Math Educ (2019) 17:405–425
https://doi.org/10.1007/s10763-017-9870-y

Mathematical Competency Demands of Assessment Items: a Search for Empirical Evidence

Andreas Pettersen[1] · **Johan Braeken**[2]

Received: 29 April 2017 / Accepted: 14 November 2017 / Published online: 8 December 2017

Abstract The implementation of mathematical competencies in school curricula requires assessment instruments to be aligned with this new view on mathematical mastery. However, there are concerns over whether existing assessments capture the wide variety of cognitive skills and abilities that constitute mathematical competence. The current study applied an explanatory item response modelling approach to investigate how teacher-rated mathematical competency demands could account for the variation in item difficulty for mathematics items from the Programme for International Student Assessment (PISA) 2012 survey and a Norwegian national grade 10 exam. The results show that the rated competency demands can explain slightly more than and less than half of the variance in item difficulty for the PISA and exam items, respectively. This provides some empirical evidence for the relevance of the mathematical competencies for solving the assessment items. The results also show that for the Norwegian exam, only two of the competencies, Reasoning and argument and Symbols and formalism, appear to influence the difficulty of the items, which questions to what extent the exam items capture the variety of cognitive skills and abilities that constitute mathematical competence. We argue that this type of empirical data from the psychometric modelling should be used to improve assessments and assessment items, as well as to inform and possibly further develop theoretical concepts of mathematical competence.

✉ Andreas Pettersen
andreas.pettersen@ils.uio.no

Johan Braeken
johan.braeken@cemo.uio.no

[1] Department of Teacher Education and School Research, University of Oslo, Blindern, P.O. Box 1099, N-0317 Oslo, Norway

[2] Centre for Educational Measurement, Faculty of Educational Sciences, University of Oslo, Blindern, P.O. Box 1161, N-0318 Oslo, Norway

Keywords Assessment items · Explanatory item response modelling · Mathematical competency demand · Sources of item difficulty

Introduction

Mathematical Competency Frameworks

In recent years, the concept of competence has gained an increased foothold in mathematics education (Boesen, Lithner & Palm, 2016; Niss, Bruder, Planas, Turner & Villa-Ochoa, 2016). While mathematics education has traditionally focused on acquiring facts and rehearsing procedures, the concept of mathematical competence covers a richer view of what it means to master mathematics (Kilpatrick, 2014; Niss et al., 2016). For instance, Niss and Højgaard (2011, p. 49) described mathematical competence as 'having the knowledge of, understanding, doing, using and having an opinion about mathematics and mathematical activity in a variety of contexts where mathematics plays or can play a role'. With this increased focus on competence in mathematics education, several competency frameworks have emerged that describe different cognitive skills and abilities that constitute mathematical competence [1] (Kilpatrick, 2014). One such example is the KOM (in Danish, 'competencies and the learning of mathematics') framework which identifies eight mathematical competencies that encompass and encapsulate the essence of what it means to master mathematics that overarchs mathematical topics and content areas (Niss & Højgaard, 2011). Kilpatrick (2014) presented concepts of mathematical competence and types of competency frameworks which have influenced curricula reforms worldwide (Boesen et al., 2016; Niss et al., 2016). For instance, the KOM framework has influenced curricula reforms in several European countries, such as Denmark, Germany, Catalonia, Sweden and Norway (Boesen, Helenius, Bergqvist, Bergqvist, Lithner, Palm & Palmberg, 2014; Niss et al., 2016; Valenta, Nosrati & Wæge, 2015).

Demands of Mathematics Assessment Items

Implementing competence-based mathematics curricula requires a shift in assessment practices, so that assessment instruments can capture the wide variety of mathematical competencies described in curriculum documents (Boesen et al., 2016; Lane, 2004; Niss, 2007). However, the assessment of competencies is regarded as challenging (Blömeke, Gustafsson & Shavelson, 2015; Koeppen, Hartig, Klieme & Leutner, 2008; Niss et al., 2016), and there are concerns over whether existing assessment items suitably assess aggregated and higher order competencies (Niss, 2007), high-level thinking skills (Lane, 2004) and complex abilities (Koeppen et al., 2008) which are important parts of mathematical competence. Recent decades have seen calls for more empirical evidence to ensure that theoretical cognitive-related constructs are actually represented in assessment items (Embretson & Gorin, 2001; Lane, 2004; Messick,

[1] In this paper the term mathematical *competence* refers to a general definition of what it means to master mathematics (e.g. the description provided by Niss and Højgaard), while the term mathematical *competency* (or *competencies* in plural) refers to one or a set of the constituent parts of mathematical competence.

1995). Messick (1995) noted that such empirical evidence can be derived from various sources. One such source is psychometric models that link theoretical cognitive models with empirical measurements (Embretson & Gorin, 2001; Koeppen et al., 2008; Messick, 1995). The modelling of item difficulty has been used as validity evidence for many tests (see e.g. Enright, Morley & Sheehan, 2002; Gorin & Embretson, 2006), and identifying features of test items that influence item difficulty is important to understand what is measured in tests (De Boeck, Cho & Wilson, 2016; Graf, Peterson, Steffen & Lawless, 2005). For instance, Embretson and Daniel (2008) scored mathematic test items for cognitive complexity on 12 variables (e.g. number of equations that had to be recalled and number of subgoals involved in the solution process) and linked these variables to item performance by using an explanatory item response model. They found that most cognitive complexity variables (such as subgoal count) were significant predictors of item difficulty (but equation recall count was not), and that these cognitive complexity variables could account for about half of the variance in item difficulty. Embretson and Daniel (2008) argued that these results supported the validity of the postulated model of cognitive complexity for mathematical problem solving.

Linking Mathematical Competency Demands and Item Difficulty

The present study uses psychometrical models to look for validity evidence by linking a theoretical competency framework to empirical measurements of mathematical competence. The study builds on the work by the PISA mathematics expert group (MEG) who developed and studied an item analysis scheme to identify the competency demands of mathematical problems. The development of the item analysis scheme was based on the concept of mathematical competence that has underpinned the PISA mathematics frameworks (Turner, Blum & Niss, 2015) and that evolved in parallel and intertwined with the KOM framework (for details of the relationship between the concept of competencies in the different frameworks, see Niss (2015)).

The MEG item analysis scheme consists of operational definitions of six mathematical competencies (see Table 1) which are a modified version of the competencies in the KOM framework where the mathematical thinking competency and the reasoning competency have been merged into Reasoning and argument, and the Aids and tools competency has been omitted (Turner et al., 2015). The scheme also includes descriptions of four different levels of demand for each competency ranging from 0 (lowest demand) to 3 (highest demand) (for the full item analysis scheme, see Turner, Blum and Niss (2015)). The results from regression analysis of PISA 2003 and PISA 2006 data showed that the rated levels of competency demands could account for a considerable amount of the variance in item difficulty (Turner, Dossey, Blum & Niss, 2013). We expand on this initial study in three ways:

The Mathematics Assessment Items

Whereas the original MEG study focused solely on 48 common items in the PISA 2003 and 2006 mathematics surveys, two different assessments were used in this study: the PISA 2012 mathematics survey (84 items administered in Norway) and the Norwegian 2014 national mathematics exam (56 items). Both assessments have been developed to

Table 1 Definitions of the six mathematical competencies in the item analysis scheme (Turner et al., 2015)

Communication. Reading and interpreting statements, questions, instructions, tasks, images and objects; imagining and understanding the situation presented and making sense of the information provided including the mathematical terms referred to; presenting and explaining one's mathematical work or reasoning.
Devising strategies. Selecting or devising a mathematical strategy to solve a problem as well as monitoring and controlling implementation of the strategy.
Mathematising. Translating an extra-mathematical situation into a mathematical model, interpreting outcomes from using a model in relation to the problem situation or validating the adequacy of the model in relation to the problem situation.
Representation. Decoding, translating between and making use of given mathematical representations in pursuit of a solution; selecting or devising representations to capture the situation or to present one's work.
Symbols and formalism. Understanding and implementing mathematical procedures and language (including symbolic expressions, arithmetic and algebraic operations), using the mathematical conventions and rules that govern them; activating and using knowledge of definitions, results, rules and formal systems.
Reasoning and argument. Drawing inferences by using logically rooted thought processes that explore and connect problem elements to form, scrutinise or justify arguments and conclusions.

measure students' mathematical competence at the end of compulsory education (i.e. at the age of 15 – 16 years), and they have been influenced by the concept of competence presented in the KOM framework (Niss, 2015; Valenta et al., 2015). However, the operationalisation of mathematical competence appears somewhat different in the two assessments. According to the PISA mathematics framework, all PISA mathematics items require students to use a range of mathematical processes and capabilities to solve problems situated in real-life contexts (Organization for Economic Co-operation and Development (OECD), 2013). A similar item format is described in the national exam's guidelines, where the items are situated in everyday contexts to test the width and depth of students' mathematical competence (Norwegian Directorate for Education and Training [Utdanningsdirektoratet], 2014). In addition, the guidelines state that many items will have a traditional format with an emphasis on procedural skills. Thus, the present study expands on Turner et al.'s (2013) study both with regard to the number and types of items included.

Teacher-Perceived Competency Demands

In the original MEG study, the PISA MEG which had developed the item analysis scheme rated the mathematical competency demands of the items. The MEG's familiarity with the PISA items and their empirical difficulty could potentially bias their ratings of competency demands and, in turn, the study results. To verify whether the results are valid even with a new set of raters who are unfamiliar with the items, we used a group of five mathematics and prospective mathematics teachers (subsequently called teachers) to establish the competency demands of the items (for further information about this group and the rating procedure, see Pettersen and Nortvedt (2017)).

Psychometric Modelling through Explanatory Item Response Modelling

In the original MEG study, estimates of the item difficulty were first computed, and then, these estimates were treated as true difficulties and used as outcome variables in a

linear regression analysis. This two-step approach ignores the estimation uncertainty and makes limited use of the actual item response data with an effective sample size of 48, the number of items. Therefore, this approach risks making some inferences unreliable. An explanatory item response model is a psychometric model that allows the cognitive demands of items to be treated as item attributes (i.e. explanatory factors) that are directly related to the success rates on the items (De Boeck et al., 2016). The psychometric model is in spirit similar to the two-step approach from the MEG study; however, it integrates all steps into one and effectively uses all available item response data (all responses of each student on each item). By building on and extending the Turner et al.'s (2013) study, the present study further explored how the rated mathematical competency demands in assessment items are related to the empirical difficulty of items. Specifically, the following core research question was addressed: To what extent do differences in teacher-rated demands of the six MEG competencies in mathematics assessment items align with the differences in the empirical item difficulty?

Method

Figure 1 shows the research design of the present study. Pettersen and Nortvedt (2017)) presented the teachers' ratings of the mathematical competency demands of the items (Fig. 1, left). The current study included student responses to the items, and it applied an explanatory item response modelling approach to investigate the relationship between the differences in the teacher-rated competency demands of the items (ΔX) and the differences in empirical item difficulty ($\Delta\beta$). Details of each assessment are given below. Then, we outline a procedure to establish the teacher-rated competency demands of the items as well as a statistical analysis approach to study the relation between competency demands and empirical item difficulties.

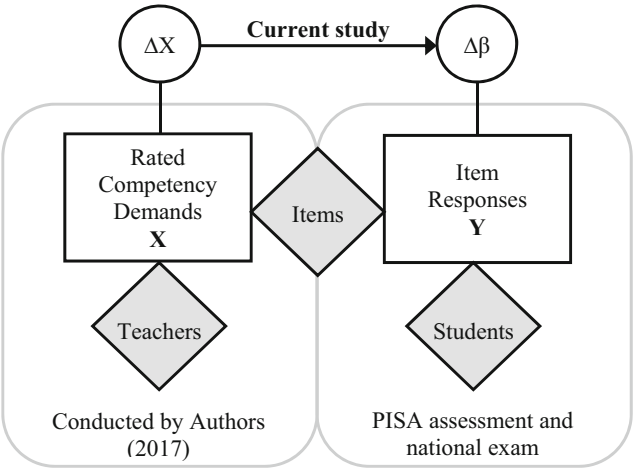


Fig. 1 Overview of study design: analysing the extent to which differences in teacher-rated competency demands (ΔX) align with differences in empirical item difficulty ($\Delta\beta$)

PISA 2012

The PISA 2012 assessment aimed to measure 15-year-old students' (i.e. students born in 1996) mathematical literacy as defined in the PISA 2012 mathematics assessment framework (OECD, 2013). According to this framework, the six mathematical competencies in the MEG item analysis scheme underpin mathematical literacy in practice (OECD, 2013). The PISA 2012 paper-based mathematics assessment was administered to a representative sample of approximately 4700 Norwegian students (OECD, 2014), and it consisted of a total of 84 items. Note that PISA is a system-level assessment, and it strategically distributes items across students by using a rotating booklet design to limit the testing time and effort (OECD, 2014). In Norway, every PISA mathematics item was assigned to between 1398 and 1452 students. In the original coding, student responses were dichotomously coded as full credit (1) and no credit (0) for 76 of the items, and eight items used full (2), partial (1) and no (0) credit codes. For ease of comparison across items, we recoded partial credits to no credit such that all student responses were dichotomously scored.

Norwegian National Exam

The Norwegian national exam in mathematics assesses students' mathematical competence based on the competence aims stated in the national curriculum. In 2014, this mathematics exam was administered to around one third of Norwegian grade 10 students; each municipality (*kommune*) was responsible for sampling its students. The present study includes a sample of 1312 students, coming from all 19 counties in Norway and from schools in both urban and rural areas. Their responses were graded by a group of specially trained markers (*oppmenn*) from different regions. The exam consisted of 56 items across two booklets (parts 1 and 2, respectively) where all students responded to both booklets. Part 1 (33 items) only allowed the use of paper, pen, ruler, compass and protractor and, according to the exam guidelines, contained both traditional non-contextualised items and contextualised mathematical problems that mainly focused on procedural skills (Norwegian Directorate for Education and Training [Utdanningsdirektoratet], 2014). Part 2 (23 items) allowed to use all types of non-communicating support material, and according to the guidelines, the items in part 2 were situated in everyday contexts and were aimed at measuring the width and depth of students' mathematical competence (Norwegian Directorate for Education and Training [Utdanningsdirektoratet], 2014). The items were originally scored by external examiners by using partial credits and subsequently rescored dichotomously such that partial and full credits were scored as 0 and 1, respectively, for ease of comparison across items.

Mathematical Competency Demands of Assessment Items

The mathematical competency demands of the assessment items, to be used as explanatory factors in the explanatory item response models, were rated by five teachers using the item analysis scheme described above. All five teachers had some experience of teaching mathematics in secondary school (ranging from 2 years of full-time teaching to a few months of part-time teaching), and they received about 1 day of training that

focused on enhancing their understanding of the competencies and the application of the item analysis scheme (Pettersen and Nortvedt (2017) have discussed this training and rating process in more detail). Overall, Pettersen and Nortvedt (2017) found that there was a high consistency in the teachers' ratings of the demands for each of the six competencies, but that higher levels were underrepresented in the ratings. In the current study, for each of the 140 assessment items, we used the average demands on each of the six competencies as rated by the teachers. The use of the teacher-averaged item ratings was supported by the high interrater consistency.

Statistical Analysis

A similar analysis procedure was followed for both assessments. First, descriptive statistics on the teacher-rated competency demands across items were given. Second, an explanatory item response modelling approach was used to study the relationship between the rated competency demands of the items and their empirical difficulty. A range of item response models with different item predictors was fitted using the lme4 package (Bates, Mächler, Bolker & Walker, 2015) in the open-source statistical software R (R Core Team, 2016) to investigate how the rated competency demands could explain item difficulties.

Item Response Models

In an item response model such as the one-parameter logistic item response model, the probability of a person p correctly responding to an item i ($Y_{pi} = 1$) is modelled as a function of the ability of the person (θ_p) and the difficulty of the item (β_i):

$$\Pr\left(Y_{pi} = 1|\theta_p, \beta_i\right) = \frac{\exp\left(\theta_p - \beta_i\right)}{1 + \exp\left(\theta_p - \beta_i\right)},$$

The person ability is estimated based on her or his performance on the test, while the item difficulty is estimated based on the performances of all persons on that item. The person abilities and item difficulties are placed on the same scale, and as can be seen from the function above, a higher person ability (θ_p) relative to the item difficulty (β_i) (either due to a more able person, an easier item or both) leads to a higher chance of a correct response. For persons with ability equal to the item difficulty ($\theta_p = \beta_i$), the chance of a correct response on that item is 50%. The explanatory extension used in this study can be seen as adding an additional regression layer to the model, $\beta_i = \sum_{k=1}^{K} X_{ik} b_k + \varepsilon_i$ (Janssen, Schepers & Peres, 2004). This extension allows the item difficulty to be predicted based on K item predictors such as the rated demands of the six mathematical competencies X_1 to X_6:

$$\Pr\left(Y_{pi} = 1|\theta_p, \beta_i\right) = \frac{\exp\left(\theta_p - \sum_{k=1}^{K} X_{ik} b_k + \varepsilon_i\right)}{1 + \exp\left(\theta_p - \sum_{k=1}^{K} X_{ik} b_k + \varepsilon_i\right)}.$$

For this explanatory model, the probability of a person p correctly responding to an item i ($Y_{pi} = 1$) is dependent on the ability of the person (θ_p) and the rated competency demands of the item (X_1 to X_6). For each of the six competencies, a regression coefficient b_k is estimated that links the demand for this competency to student performance. As increased ratings are to reflect more demanding items, we expected that higher ratings would be related to more difficult items, and thus that the estimated regression coefficients should be positive (i.e. $b_k>0$) for all competencies. The extent to which the item difficulties can be replaced by the teacher-rated competency demands and their estimated regressions coefficients was used to investigate to what extent the demand for the six MEG competencies could explain the variation in item difficulties. Formally, this model is equivalent to a multilevel cross-classified logistic regression model with responses nested in students and nested in items, with both student ability and item difficulty modelled as a random effect. In the field of mathematics education, Embretson and Daniel (2008) have already used a reduced form of such a model to understand and quantify the cognitive complexity in mathematical problem solving items.

Missing Data

For both assessments, missing responses were coded as 0 if students had attempted to solve at least one of the subsequent items. Consecutive missing responses clustered at the end of the tests were regarded as not reached and subsequently omitted, except for the first value in the missing series, which was coded as 0. This treatment of missing data is common in large-scale assessments (e.g. in PISA, OECD, 2014), and it is performed to avoid inflating the item difficulty due to confounding by the time constraints of the test and working speed of students. For the Norwegian exam, the missing data in the two parts were treated separately, as they each had an individual time frame.

Results

PISA 2012

Teacher-Rated Competency Demands in Mathematics Assessment Items

Table 3 shows the descriptive statistics on the five teachers' ratings of the mathematical competency demands of the PISA items. The intraclass correlation coefficients (ICC) indicated that the teachers had a rather high agreement in their ratings of each of the six competency demands. The means of the rated competency demands were between 0.56 and 1.22 for the PISA items. The distribution of the averaged teacher ratings are displayed in Fig. 4 and shows that Communication and Reasoning and argumentation were the highest-rated competencies in the PISA items and that most ratings for all competencies were located at lower levels (levels 0 and 1). The ratings of competency demands tended to be moderately positively intercorrelated (median $r = .39$), indicating some overlap in the teacher-rated competency demands across the different competencies (see Table 3). Although the correlation between Communication and

Devising strategies was somewhat lower (r = .19), the real exception to the rule was Representation, the ratings for which were much less correlated with the ratings for the five other competencies (median r = − .08).

Relationships Between Rated Competency Demands and Empirical Item Difficulty

Null Model The average 15-year-old Norwegian student has 42% chance of correctly responding to an average PISA 2012 mathematics item. It was found that 21 and 35% of the variation in response was attributable to individual student differences and individual item differences, respectively. Therefore, although there is quite some variation in student ability (σ^2_θ = 1.61), there is even more variation in the item difficulty (σ^2_β = 2.68) for PISA 2012; these latter differences in item difficulty we hypothesised to be linked to differences in rated mathematical competency demands.

Single-Predictor Models First, we assessed the total impact of each competency demand separately by fitting six explanatory item response models with one item predictor each—a single competency demand—to explain the variation in item difficulty (see Table 4). Likelihood ratio tests showed that each of the six explanatory models had significantly better fit than the null model, except for the Representation model. This was corroborated by the relative fit model comparisons; the Representation model was the only one with Akaike's information criterion (AIC)[2] larger than that of the null model. For most other competencies, the differences in teacher-rated competency demands accounted for around 30% of the variation in item difficulty. This corresponds to rather large effects of the differences in rated competency demands. For instance, a 1-point rating increase in the demand for Symbolism and formalism goes together with a 5.61 (i.e. exp.($b_{1(k)}$)) multiplicative decrease in the odds of correctly responding to the item. In terms of probability, this would mean that an average student would have a 70% chance of responding correctly to an item with a rated demand of 0 for Symbolism and formalism, whereas this chance would decrease to 29% for an item rated at level 1.

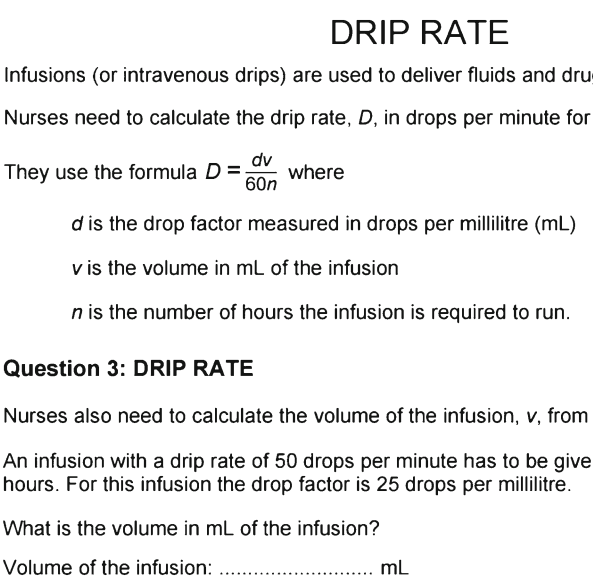
Full Model In the second stage, we assessed the impact of the mathematical competency demands in their full context by fitting an explanatory item response model with the rated demands for all six competencies as item predictors to explain the variation in item difficulty (see Table 4). This full model outperformed the null model and each single-predictor model, indicating that taking into account the full item competency demands profile improves the prediction of item difficulty. Around 55% of the variance in item difficulty could be explained by variation in competency demands; this represents a 25% increase compared to the explanatory power of a single competency demand. In line with expectations, the effect of each competency demand was positive, meaning that competency demands increase with increased item difficulty. The strongest predictors for item difficulty in the full model were Reasoning and argument, Symbols and formalism and Devising strategies. Surprisingly, upon keeping the other competency demand ratings constant, a one level increase in rated demand for

[2] AIC Akaike's information criterion (Akaike, 1973) balances absolute fit to the data with model complexity in terms of the number of parameters. Best model is a model that is parsimonious but still fits well.

Representation related to an increase of half a point in item difficulty on the logit scale. Therefore, while the demand for Representation on its own is not informative (cf. single-predictor model), it becomes relevant once it is seen in the context provided by the other competency demands. Conversely, the contributions of Communication and Mathematising were no longer significantly different from zero in the context of the full competency demands profile due to the multicollinearity with these other competencies.

Figures 2 and 3 show two of the PISA mathematics items included in the current study. Table 2 shows the averaged teacher-rated mathematical competency demands, the predicted difficulty (based on the rated competency demands and regression coefficients $[1.8 \times .5 + 0.8 \times .73 + 0.8 \times .01 + 1.0 \times .49 + 2.2 \times .88 + 1.2 \times .99 - 2.81 = 2.30]$) and the empirical difficulty of the two items. For 'drip rate', the predicted and empirical item difficulties are rather equal (see Table 2). This indicates that for this item, the model adequately captures the difficulty of the item based on its rated competency demands. For 'sauce', we observe that the empirical difficulty is lower than the predicted difficulty (Table 2). This means that the item was less difficult for the students than what was expected based on the rated competency demands of the item.

Holistic Models In the third stage, we assessed the explanatory power of the simple holistic summary scores of the six competency demand ratings. With 21% of explained variance in item difficulty, using the number of competency demands rated above 1 did not prove fruitful for providing a practical summary that reflects the item difficulty (see Table 4). However, this percentage increases to 41% when using the number of competency demands rated above 0 as predictor. With 53% of explained variance in item difficulty, summing up the rated demands across the six competencies provides a sum score that matches the predictive performance of the full model (55%). As supported by the AIC model comparisons, the sum-across-competencies model is an


DRIP RATE

Infusions (or intravenous drips) are used to deliver fluids and drugs to patients.

Nurses need to calculate the drip rate, D, in drops per minute for infusions.

They use the formula $D = \frac{dv}{60n}$ where

d is the drop factor measured in drops per millilitre (mL)

v is the volume in mL of the infusion

n is the number of hours the infusion is required to run.

Question 3: DRIP RATE

Nurses also need to calculate the volume of the infusion, v, from the drip rate, D.

An infusion with a drip rate of 50 drops per minute has to be given to a patient for 3 hours. For this infusion the drop factor is 25 drops per millilitre.

What is the volume in mL of the infusion?

Volume of the infusion: ..... mL

Fig. 2 PISA 2012 mathematics item 'drip rate'

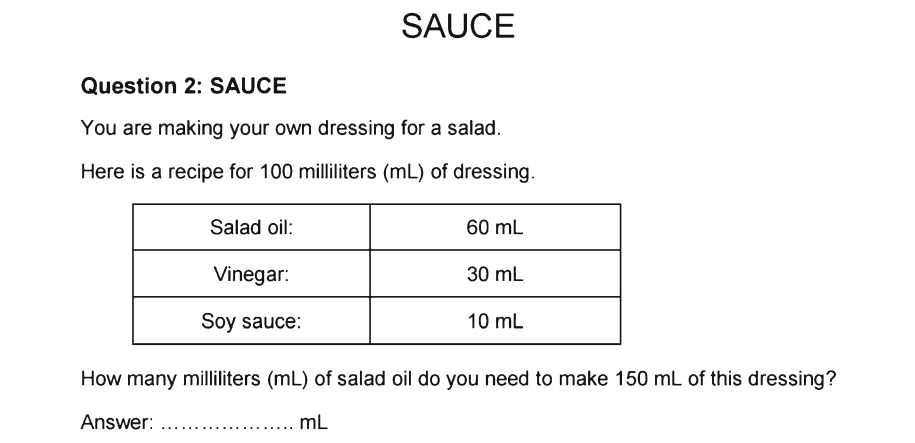


Fig. 3 PISA 2012 mathematics item 'sauce'

equally well fitting, but more parsimonious model than the full model. Therefore, from a practical viewpoint, the sum of the mathematical competency demands might be a good and intuitive indicator for teachers to estimate the difficulty of items similar to those in the PISA assessment. For every point increase in demands, the odds of correctly responding to the item decreases by a factor of 2 (i.e. exp.(.61) = 1.84). In terms of probability, this means that an average student is predicted to have a probability of correctly responding to around 93, 81, 57 and 29% for an item with total competency demand ratings of 0, 2, 4 and 6, respectively.

Norwegian National Exam

Teacher-Rated Competency Demands in Mathematics Assessment Items

Table 3 shows descriptive statistics on the teachers' ratings of the mathematical competency demands of the exam items. High agreement was observed in the teachers' ratings of all six competencies, as indicated by the ICCs. As seen from the distribution of the teachers' ratings in Fig. 4, Symbols and formalism received somewhat higher ratings than the other competencies, and Representation received the lowest ratings. As reflected by the median correlation of $r = .51$, the rated competency demands of the exam items tended to be moderately positively intercorrelated (see Table 3), indicating

Table 2 Averaged teacher-rated competency demands, predicted item difficulty and empirical item difficulty of the two example items 'drip rate' (Fig. 2) and 'sauce' (Fig. 3)

Item	Teacher-rated competency demands						Predicted item difficulty	Empirical item difficulty
	C	DS	M	R	SF	RA		
Drip rate	1.8	0.8	0.8	1.0	2.2	1.2	2.30	2.31
Sauce	0.6	1.0	0.2	0.4	0.6	1.2	0.13	− 0.80

C communication, *DS* devising strategies, *M* mathematising, *R* representation, *SF* symbols and formalism, *RA* reasoning and argument

Table 3 Descriptive statistics of the five teachers' competency demand ratings: agreement measures (ICC), means, standard deviations (SD) and correlations between competencies

	PISA items								Exam items							
				Correlations								Correlations				
	ICC	Mean	SD	1.	2.	3.	4.	5.	ICC	Mean	SD	1.	2.	3.	4.	5.
1. Communication	.77	1.22	.48						.89	.76	.71					
2. Devising strategies	.86	.77	.59	.19					.92	.63	.69	.70				
3. Mathematising	.77	.56	.49	.41	.62				.84	.40	.49	.51	.73			
4. Representation	.84	.58	.51	.11	− .24	− .08			.79	.27	.36	.19	.06	− .11		
5. Symbols and formalism	.83	.67	.55	.39	.62	.60	− .22		.82	1.10	.52	.45	.51	. 36	− .15	
6. Reasoning and argument	.74	1.11	.48	.40	.42	.58	.01	.32	.89	.65	.62	.60	.67	.85	.02	.26

ICC values are reproduced from Pettersen and Nortvedt (2017). Means, standard deviations and correlations are based on the average ratings of the five teachers and prospective teachers

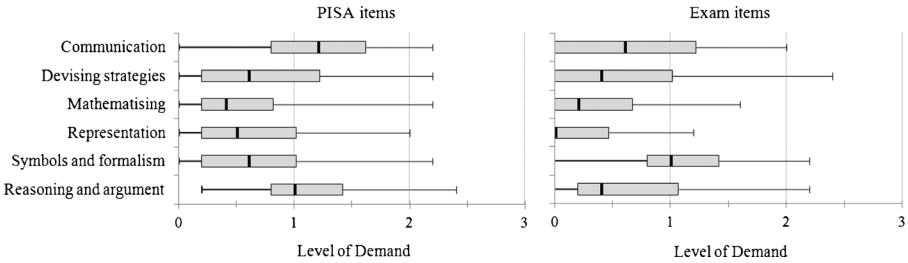


Fig. 4 Boxplots of the distribution of the averaged teacher ratings of mathematical competency demands

a substantial overlap in rated demands across the different competencies. Devising strategies correlated highest with the other competencies (median r = .67), Symbols and formalism intercorrelated to a lower degree (median r = .36), and Representation appeared to be uncorrelated to the other five competencies (median r = .02).

Relationships Between Rated Competency Demands and Empirical Item Difficulty

Null Model The average Norwegian grade 10 student has around 54% chance of correctly solving an item of average difficulty on the 2014 national mathematics exam. It was found that 31 and 33% of the variation in responses was attributable to individual student differences and individual item differences, respectively. Therefore, the levels of variation in student ability (σ^2_θ = 2.85) and item difficulty (σ^2_β = 3.08) are rather similar; these latter differences in item difficulty we hypothesised to be linked to differences in mathematical competency demands.

Single-Predictor Models Six explanatory item response models, each using one competency demand as an item predictor, were fitted to examine the extent to which a single competency demand on its own could explain the variation in item difficulty (see Table 4). According to both the AIC values (see Table 4) and Chi-square likelihood ratio tests, all explanatory item response models showed significantly better fit than the null model, except for the Representation demand model. For the five other single-predictor models, the differences in the teacher-rated demands for each competency accounted for 13 − 28% of the variation in item difficulty. For instance, a 1-point rating increase in the demand for Symbolism and formalism goes together with a 5.40 (i.e. exp.($b_{1(k)}$)) multiplicative decrease in the odds of correctly responding to the item. In terms of probability, this would mean that an average student would have an 88% chance of responding correctly to an item rated at level 0 for Symbolism and formalism, whereas this chance would decrease to 58 and 21% for an item rated at levels 1 and 2, respectively.

Full Model An explanatory item response model with the demand of all six competencies as item predictors was fitted to explain the variation in item difficulty (see Table 4). According to the Chi-square likelihood ratio test and AIC, this full model outperformed both the null model and the single-predictor models. In total, the rated demands of all six competencies accounted for 48% of the variance in item difficulty. This is an increase of 20% from the best-performing single competency model.

Table 4 Parameter estimates and model fit for the explanatory item response models of the two assessments based on the teachers' averaged ratings of competency demands and student responses

	PISA data					Exam data				
	Full model		Single-predictor models			Full model		Single-predictor models		
Predictors	b_k	SE	$b_{1(k)}$	$R(b)^2$ (%)	AIC	b_k	SE	$b_{1(k)}$	$R(b)^2$ (%)	AIC
(Intercept)	− 2.81**	.42			116163[a]	− 2.78**	.49			62959[a]
Communication	.50	.31	1.54	20	116,146	− .25	.39	.91	13	62,953
Devising strategies	.73*	.30	1.56	32	116,133	− .37	.47	.93	13	62,953
Mathematising	.01	.39	1.87	31	116,133	− 1.30	.83	1.47	16	62,951
Representation	.49*	.25	.14	0	116,165	.04	.54	− .21	0	62,961
Symbols and formalism	.88**	.33	1.72	33	116,131	1.78**	.42	1.69	24	62,946
Reasoning and argument	.99**	.32	1.89	31	116,134	2.46**	.61	1.52	28	62,943
Full model R^2				55	116,108				48	62,935
			Holistic models					Holistic models		
Sum demand across competencies			.61	53	116,101			.36	26	62,944
Number of competencies rated > 0			.69	41	116,121			.52	21	62,948
Number of competencies rated > 1			1.30	21	116,145			1.17	15	62,952

b regression coefficients, *SE* standard errors, $R(b)^2$ % explained variance in item difficulty, *AIC* Akaike's information criterion

**$p < .01$, *$p < .05$

[a] Fit values for null model

Reasoning and argument and Symbols and formalism were the two prominent competency demands in the full model that seemed to account for most of the explained variation in item difficulty. The partial effects of Reasoning and argument and Symbols and formalism (i.e. keeping the value of the other predictors constant) correspond to a multiplicative decrease in the odds of correctly responding to the item of around 12 and 6, respectively (i.e. exp.(2.46) and exp.(1.78)). The partial effects of the other competencies were not statistically significant from zero, and hence, their unique impact disappeared in the presence of these two strong item predictors.

Holistic Models The explanatory power of three models that represented a more holistic view of the demand of the items was assessed. The AIC values in Table 4 show that all three models have a worse fit than the full model. Furthermore, the explanatory powers of the holistic models were rather low compared to that of the full model, and also lower than that of the best-fitting single-predictor model. Of the three holistic models, the sum of the demand across all six competencies had the most explanatory power, accounting for 26% of the variance in item difficulty. The low predictive power of these models compared to that of the full model is in line with the observed coefficient imbalance in the full model, where Reasoning and argument and Symbols and formalism dominate the predictive equation. Combining the latter, two competency demand ratings constitute a good summary predictor of item demands in

the Norwegian exam; adding up the remaining competency demand ratings would only add noise to the signal.

Discussion

The implementation of competence-based mathematics curricula requires assessment instruments that ensure valid measures of mathematical competence (Boesen et al., 2016; Lane, 2004; Niss & Højgaard, 2011). The present study aimed to investigate the relationship between rated mathematical competency demands of assessment items and item difficulty in a search for empirical evidence that indicates whether the items draw on the six MEG competencies.

Teacher-Rated Competency Demands in Mathematics Assessment Items

According to the teachers' ratings of mathematical competency demands, both the PISA and Norwegian exam items require the activation of all six competencies to some extent, although the vast majority of the demands were rated at lower levels of 0 and 1. For the PISA items, Communication and Reasoning and argument were perceived as the two prominent competencies, and Symbols and formalism was the far most dominant competency in the teachers' ratings of the Norwegian exam items. For both assessments, the demands for Representation, and to some extent Mathematising, appeared almost absent in the teachers' ratings. The dominant presence of Symbols and formalism indicates a focus on procedural knowledge, mathematical rules and definitions (e.g. calculations, algebraic operations and formulas) and resembles the traditional viewpoint of mathematics curriculum and assessments.

Overall, a moderate intercorrelation was observed in the rated demands for the different competencies. Because of the overlapping nature of the mathematical competencies, it was expected that some correlations would occur between them. Therefore, it was uncertain whether the moderate correlations could be due to the defined overlaps of the competencies or the teachers' inability to distinguish between the demands for the different competencies due to unclear definitions and operationalisations in the item analysis scheme. The substantial correlation observed between Devising strategies and Mathematising could indicate that the two competencies tended to be demanded in pair due to the design of the items, or that too vague distinctions caused the teachers to lump together the demands for these competencies. At the same time, the low correlation between Representation and other competencies could indicate the clearer distinction of the demand for Representation. A clear distinction of this competency combined with its low presence in the teachers' ratings raises the question of whether the demand for this competency is actually more or less absent in the items or whether the operationalisation of Representation in the item analysis scheme does not capture the demand for this competency adequately.

Based on the rather high agreement in the teachers' ratings, we assume that the overall moderate intercorrelations in the rated competency demands are mainly driven by the defined overlaps of the competencies and that the teachers' average ratings provide reliable ratings of the competency demands of the assessment items.

Relationship Between Competency Demands and Empirical Item Difficulty

The different explanatory item response models provided complimentary information about the relationship between the rated mathematical competency demands and item difficulty. This information could be used to further verify the teacher-rated competency demands and to investigate the extent to which the items actually seem to tap into the six mathematical competencies in the MEG framework. The single-predictor explanatory item response models show that, on their own, most teacher-rated competency demands could explain a substantial part of the variation in item difficulty in both assessments. The relationship between competency demands and item difficulty were most prominent for Reasoning and argument and Symbols and formalism, whereas the rated demand for Representation turned out to be largely irrelevant on its own. The latter is a surprising finding as being able to interpret, translate between and devise different types of mathematical representations is considered important for solving mathematical problems (Duval, 2006; Elia, Panaoura, Eracleous & Gagatsis, 2007; National Council of Teachers of Mathematics (NCTM), 2000; Stylianou, 2011). However, this result does need further scrutiny; as according to the teachers' ratings, the two assessments have low to almost absent demands for the Representation competency.

The inclusion of the rated demands for all six mathematical competencies yielded better models of the item difficulties than the single-predictor models. The full explanatory item response models could explain about half of the variance in item difficulty (55% for PISA and 48% for the Norwegian exam), indicating that the competency demands profile of an item would be a decent indicator of that item's difficulty. When examining the effects of the different competencies in the full models, the rated demands for Symbols and formalism and Reasoning and argument appeared to be the most strongly related to the item difficulty in both assessments. For the exam items, the demands for these two competencies were the only ones related to item difficulty; the difficulty of the PISA items was also related to the rated demands for Devising strategies and Representation. Therefore, although, on its own, the Representation demand did not make a difference; it became relevant when taking into account the context provided by the demands of the other competencies in the PISA items. The results based on the PISA 2012 data in the present study are consistent with Turner et al.'s (2013) findings based on PISA 2003 and PISA 2006 data, except for the relevance of the Representation demand. In our study, the Representation effect does appear relatively stable even when removing PISA items with the highest Representation ratings. Therefore, we conjecture that Representation demands are not strong enough on their own (i.e. single predictor) but can only play a role in the context provided by the demand for other competencies (i.e. multiple predictors). For instance, as symbolic representations play an important role in mathematics, an item that requires decoding symbolic representations (Representation competency) would often also require the use of these symbolic representations in formal operations such as solving equations (Symbols and formalism competency).

The usefulness of a more overall competence perspective was explored in the holistic models. The holistic perspective was unsuitable for the exam items, as aggregating ratings would lead to mostly adding noise to the signal of the two dominant predictors, Symbols and formalism and Reasoning and argument. For the PISA items,

the holistic perspective proved more successful due to the more equal distribution of explanatory power across the competency demands.

Both the PISA assessment and the Norwegian exam aimed to measure general mathematical competence, and the PISA mathematics framework and Norwegian curriculum are both influenced by the concept of mathematical competence described in the KOM framework (Niss, 2015; Valenta et al., 2015). Therefore, even though none of the assessments was designed to explicitly measure the six mathematical competencies in the MEG framework, we would expect that these competencies should be relevant to the constructs of mathematical competence represented in both PISA and the exam. The descriptions of the assessment items in the PISA 2012 mathematics framework and the 2014 Norwegian exam guidelines indicate somewhat different operationalisations of mathematical competence in the two assessments. While all PISA items are situated in real-life contexts developed to measure a wide range of processes and capabilities (OECD, 2013), the Norwegian exam consists of both contextualised items aimed at measuring the depth and width of students' mathematical competence and non-contextualised items focused on procedural skills (Norwegian Directorate for Education and Training [Utdanningsdirektoratet], 2014). The results from the explanatory item response models indicate that the MEG competency framework seems to be better reflected in the PISA items than in the exam items, with more of the variance in item difficulty being explained and a higher number of competencies related to item difficulty. The fact that the rated demands for only two of the six competencies could be related to the difficulty of the exam items raises questions about the extent to which these items require various cognitive skills and abilities that are represented in mathematical competence. These results could indicate that the narrow focus on procedural skills for a rather large proportion of the exam items might be at the expense of a valid measure of more general mathematical competence as represented through the six MEG competencies.

Implications for MEG Competency Framework and Assessment of Competency Demands

The present study provides some promising results for applying an explanatory item response modelling approach to link mathematical competency demands to item difficulty. The models show that a substantial amount of the variance in item difficulty can be explained by the teacher-rated competency demands, indicating the relevance of the mathematical competencies for solving the assessment items. Nonetheless, about half of the variance in item difficulty is not explained by the rated competency demands. A part of this unexplained variance could be related to inaccurate ratings of the mathematical competency demands and difficulties with distinguishing the demand for different competencies. More rater training and rater experience and clearly defined and concrete categories have been recognised as factors that can increase the discrimination between distinct concepts and dimensions (Feeley, 2002). Thus, further revisions of the scheme (e.g. clarifying definitions and descriptions) and guiding material (e.g. more items to exemplify differences between the competencies and levels of demand) could yield more accurate ratings that better reflect the empirical difficulty of the items. In addition, a more exhaustive training (for instance by expanding from one to two full days of training) where more time was spent on applying the scheme,

comparing ratings and discussing differences could improve the raters' ability to distinguish between several levels of demand. Nonetheless, it is likely that most of the unexplained variance is related to item features that are not related to the six mathematical competencies addressed in our study. Identifying features that influence the item difficulty is important for understanding what is being measured in assessments and for ensuring that construct-irrelevant item attributes do not threaten the validity of the interpretations of the test scores (De Boeck et al., 2016; Graf et al., 2005). Although the variance in item difficulty cannot be fully accounted for by any set of item features, it is probable that item features exist that are relevant to mathematical competence but not captured in the MEG scheme. Turner et al. (2013, p. 24) noted that the mathematical competencies included in the item analysis scheme 'describe the essential activities when solving mathematical problems' and were based on a reconfiguration of the KOM competencies. One obvious difference between the KOM framework and the MEG framework is that the Aids and tools competency, which was not relevant for the early paper-based PISA items, is not a part of the MEG competencies. In part 2 of the Norwegian exam, the students are required to use digital tools (e.g. spreadsheets) for solving some of the items (Norwegian Directorate for Education and Training [Utdanningsdirektoratet], 2014). As this activity is not captured by the teachers' ratings, the demands for aids and tools might explain some of the variances in student performance in the exam data. However, by inspecting residuals between the empirical and the estimated item difficulties based on the modelled competency demands, exam items that require the use of digital tools do not appear to be more divergent than the other items. This indicates that the inclusion of an aids and tools demand in the item response models would likely not influence the results to a large extent.

When investigating the empirical separability of cognitive and content domains in mathematical competencies, Harks, Klieme, Hartig and Leiss (2014) found that the competencies included in their study were content-specific; the level of demand for the competencies depended on the mathematical content in which they appear. One characteristic of the KOM and MEG competency frameworks that distinguishes them from many other mathematical frameworks is the absence of content domains (e.g. algebra, geometry and measurement). Previous studies that have investigated features that influence the difficulty of mathematics items are mostly situated within a certain topic area. For instance, both Enright et al. (2002) and Koedinger and Nathan (2004) investigated the factors that influenced the difficulty of word problems related to quantity. Although the competencies in the MEG framework are defined to overarch mathematical topics and content, their activation could be more or less demanding within certain domains or areas. It is therefore possible that including information about content or context as item attributes could account for some of the unexplained variances.

Wilson, De Boeck and Carstensen (2008) distinguished between planned and unplanned variation in item properties, where the former refers to items that are developed to systematically vary with regard to the properties of interest, and the latter is based on post hoc analysis of items in an extant test. Studies using planned variation have shown that a large proportion of the variance in item difficulty can be explained by the varied item features (Daniel & Embretson, 2010; Enright et al., 2002). In the case of PISA and the Norwegian exam, the variation in mathematical competency

demands is planned on the conceptual level, whereas the implementation has been much more ad hoc with a more indirect impact on the development of the assessment items. Therefore, a next step to follow-up on this research is to replicate the study with a more systematically designed assessment that follows through the competency framework from concept to implemented items. The latter might prove to be the real challenge. One might wonder whether it is practically feasible to design item sets that tap into each combination of both competencies and levels of demand or that isolate the demand of different competencies.

Concluding Remarks

Zlatkin-Troitschanskaia, Shavelson and Kuhn (2015) stress that the conceptual model of competency should dictate the nature of the psychometric models and not the other way around. Although only a subsample of the six mathematical competencies is identified empirically and seems to add to the explanatory power in the item response models, reducing the theoretical competency framework to include only a subsample of competencies would be an improper interpretation and use of empirical data. For instance, the importance of communication and representations is recognised in mathematics frameworks, curricula and assessments around the world (Niss et al., 2016), and removing these would lead to a limited view of mathematical competence. Rather, empirical data from psychometric modelling should be used to inform and further develop theoretical models of mathematical competence, as well as to improve assessments and operationalisation of competencies in assessment items. From this perspective, we call not only for the continued conceptual development of the competency framework but also for a more systematic development of assessment items that are intended to tap into these competencies. The latter area has been neglected for too long and is vital to ensure valid measures of mathematical competencies that are aligned with the present goals of mathematics education.

Acknowledgements The authors would like to thank Ross Turner for his support, feedback and contribution of material during this study. Further, we would also like to thank the teachers and prospective teachers for their contribution, the Norwegian PISA Group for allowing access to the PISA material and the Norwegian Directorate for Education and Training for access to the national exam material.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Aademiai Kiado.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies: Viewing competence as a continuum. *Zeitschrift für Psychologie, 223*(1), 3–13.

Boesen, J., Helenius, O., Bergqvist, E., Bergqvist, T., Lithner, J., Palm, T., & Palmberg, B. (2014). Developing mathematical competence: From the intended to the enacted curriculum. *The Journal of Mathematical Behavior, 33*, 72–87.

Boesen, J., Lithner, J. & Palm, T. (2016). Assessing mathematical competencies: An analysis of Swedish national mathematics tests. *Scandinavian Journal of Educational Research*, 1–16. https://doi.org/10.1080/00313831.2016.1212256.

Daniel, R. C., & Embretson, S. E. (2010). Designing cognitive complexity in mathematical problem-solving items. *Applied Psychological Measurement, 34*(5), 348–364.

De Boeck, P., Cho, S. J., & Wilson, M. (2016). Explanatory item response models. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 249–268). Hoboken: Wiley.

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

Elia, I., Panaoura, A., Eracleous, A., & Gagatsis, A. (2007). Relations between secondary pupils' conceptions about functions and problem solving in different representations. *International Journal of Science and Mathematics Education, 5*(3), 533–556.

Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science, 50*(3), 328–344.

Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343–368.

Enright, M. K., Morley, M., & Sheehan, K. M. (2002). Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education, 15*(1), 49–74.

Feeley, T. H. (2002). Comment on halo effects in rating and evaluation research. *Human Communication Research, 28*(4), 578–586. https://doi.org/10.1111/j.1468-2958.2002.tb00825.x.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement, 30*(5), 394–411.

Graf, E. A., Peterson, S., Steffen, M., & Lawless, R. (2005). *Psychometric and cognitive analysis as a basis for the design and revision of quantitative item models* (No. RR-05-25). Princeton: Educational Testing Service.

Harks, B., Klieme, E., Hartig, J., & Leiss, D. (2014). Separating cognitive and content domains in mathematical competence. *Educational Assessment, 19*(4), 243–266.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 189–212). New York: Springer.

Kilpatrick, J. (2014). Competency frameworks in mathematics education. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 85–87). Dordrecht, The Netherlands: Springer.

Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *The Journal of the Learning Sciences, 13*(2), 129–164.

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Zeitschrift für Psychologie, 216*(2), 61–73.

Lane, S. (2004). Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice, 23*(3), 6–14.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics*. Reston: NCTM.

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, NC: Information Age.

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

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., & Højgaard, T. (Eds.). (2011). *Competencies and mathematical learning*. Denmark: Roskilde University.

Norwegian Directorate for Education and Training [Utdanningsdirektoratet]. (2014). *Eksamensveiledning - om vurdering av eksamensbesvarelser. MAT0010 Matematikk. Sentralt gitt skriftlig eksamen. Grunnskole* [Manual - to be used to assess exam papers. MAT0010 Mathematics. National written exam, end of compulsory education]. Oslo: Utdanningsdirektoratet.

Organization for Economic Co-operation and Development (OECD). (2013). *PISA 2012 Assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.* Paris: OECD Publishing. https://doi.org/10.1787/9789264190511-en.

Organization for Economic Co-operation and Development (OECD). (2014). *PISA 2012 technical report*. Paris: OECD Publishing. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Pettersen, A., & Nortvedt, G. A. (2017). Identifying competency demands in mathematical tasks: recognising what matters. *International Journal of Science and Mathematics Education*. https://doi.org/10.1007/s10763-017-9807-5.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Stylianou, D. A. (2011). An examination of middle school students' representation practices in mathematical problem solving through the lens of expert work: Towards an organizing scheme. *Educational Studies in Mathematics, 76*(3), 265–280.

Turner, R., Blum, W., & Niss, M. (2015). Using competencies to explain mathematical item demand: A work in progress. In K. Stacey & R. Turner (Eds.), *Assessing mathematical literacy: The PISA experience* (pp. 85–115). New York: Springer.

Turner, R., Dossey, J., Blum, W., & Niss, M. (2013). Using mathematical competencies to predict item difficulty in PISA: A MEG study. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.), *Research on PISA* (pp. 23–37). New York: Springer.

Valenta, A., Nosrati, M., & Wæge, K. (2015). *Skisse av den «ideelle læreplan i matematikk»* [Draft of the «ideal curriculum in mathematics»]. Trondheim: Nasjonalt senter for matematikk i opplæringen. Retrieved from https://nettsteder.regjeringen.no/fremtidensskole/files/2014/05/Skisse-av-den-ideellel%C3%A6replanen-i-matematikk.pdf.

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In E. Klieme & D. Leutner (Eds.), *Assessment of competencies in educational contexts: State of the art and future prospects* (pp. 91–120). Göttingen: Hogrefe & Huber.

Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Kuhn, C. (2015). The international state of research on measurement of competency in higher education. *Studies in Higher Education, 40*(3), 393–411.