



# Identifying Competency Demands in Mathematical Tasks: Recognising What Matters

Andreas Pettersen<sup>1</sup>  · Guri A. Nortvedt<sup>1</sup>

Received: 9 September 2016 / Accepted: 13 February 2017 / Published online: 11 March 2017  
© Ministry of Science and Technology, Taiwan 2017

**Abstract** In more and more countries, the goal of mathematics education is to develop students' mathematical competence beyond procedural and conceptual knowledge. Thus, students need to engage in a rich variety of tasks comprising competencies such as communication, reasoning and problem-solving. In addition, assessment tasks should be constructed to measure the particular aspects of mathematical competence to align assessment with curriculum. Teachers need to be able to recognise competency demand in the mathematical tasks they want to use for teaching and assessment purposes, which might prove difficult. The aim of this study was to analyse the outcome of 5 teachers' and prospective teachers' use of an item analysis tool. For each of 141 assessment tasks, the teachers and prospective teachers individually applied the tool to identify the competency demand of the task on a scale from 0 to 3 for each of 6 mathematical competencies. Overall, the analysis reveals high consistency in their analysis. However, the teachers and prospective teachers utilised a restricted range of the scale, rarely judging a task to demand a high level of competence. This indicates that the 5 teachers and prospective teachers can use the tool to identify which of the 6 competencies are at play in solving a task, but can only differentiate to a limited extent between tasks that demand a low level of competence and those that demand a high level. In conclusion, we propose that an item analysis tool could be useful to teachers and prospective teachers as a means of analysing and selecting appropriate tasks that enhance development of mathematical competencies.

---

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

Guri A. Nortvedt  
guri.nortvedt@ils.uio.no

<sup>1</sup> Department of Teacher Education and School Research, University of Oslo, P.O. Box 1099, Blindern, 0317 Oslo, Norway

**Keywords** Mathematical competence · Teachers' task analysis · Task complexity · Cognitive demand in tasks · Task analysis tool

## Introduction

In recent decades, the increased focus on competence has influenced curriculum design, leading to the introduction of concepts like 'competence learning' and 'competence-based curricula' (Koeppen, Hartig, Klieme & Leutner, 2008; Westera, 2001). Consequently, current curricula often comprise a rich view of what it is to know mathematics by focusing on competencies and including aspects of mathematical literacy (Burkhardt, 2014). The movement towards more competence-based curricula requires changes in teacher competence and teaching practices. Across the world, mathematical tasks play a central role in enhancing students' learning and understanding (Shimizu, Kaur, Huang & Clarke, 2010). Change in classroom practices necessitates a change in the kind of tasks offered in teaching and learning situations towards tasks that can foster development of mathematical competencies. However, such change is not necessarily obtained through changes and reforms in curricula documents alone (Boesen et al., 2014). Prior research demonstrates that the tasks offered in mathematics classrooms have been and still are quite uniform, with a strong focus on developing procedural knowledge. For example, in the TIMSS 1999 Video Study, for five of the six countries where data were available, on average, more than half of the problems presented for a lesson focused on fostering procedural skills rather than making connections or stating concepts (Hiebert et al., 2003). More recently, Kaur (2010) found that practice and assessment tasks in Singapore classrooms focus mainly on facts and procedures; similarly, studies indicate that recalling algorithms and using procedures are the main focus of both textbook exercises (Lithner, 2004) and teacher-made tests (Palm, Boesen & Lithner, 2011) in Sweden.

To select appropriate tasks to extend and assess students' knowledge, teachers must be able to analyse the demand of the tasks with respect to both content and cognitive demand (Shimizu et al., 2010). Arbaugh and Brown (2005) demonstrated that this could be challenging and that when analysing tasks, teachers tend to focus mainly on surface features rather than on in-depth features. The aim of the present study was to analyse the outcome of a group of five teachers' and prospective teachers' use of an item analysis scheme for identifying the competency demands of assessment tasks ( $n = 141$ ). The research question regards the degree to which the group of teachers and prospective teachers consistently analyse the competency demands of tasks originally developed to assess students' mathematical competence. Consequently, tasks from a Norwegian national exam and from the Programme for International Student Assessment (PISA) 2012 study were selected in order to provide the teachers and prospective teachers with a large variety of tasks comprising different levels of cognitive demand.

## Changing What Matters: towards a Competence-Based Mathematics Curriculum

Although the term ‘competence’ lacks a common definition and understanding (Blömeke, Gustafsson & Shavelson, 2015; Pikkariainen, 2014; Westera, 2001), it is usually characterised as the ability to apply knowledge and skills in order to master complex situations (Westera, 2001). The concept of competence is often regarded as context-specific, which means that the acquisition of competencies is based on “learning and experience in relevant, domain-specific situations” (Koeppen et al., 2008, p. 62). The concept of mathematical competence is not a novel idea. In 1986, Hiebert was already arguing that for students to be fully competent in mathematics, they need both conceptual and procedural knowledge and an understanding of the relationship between the two. Hiebert (1986) was commenting on the long-standing tradition in mathematics education in which conceptual and procedural knowledge were viewed as separate entities, and where researchers tended to debate which of these kinds of knowledge students should be taught first. More recently, a richer view of mathematics and mathematical competence has evolved. By the early 2000s, several frameworks had emerged emphasising not only the interaction between conceptual and procedural knowledge but also the importance of abilities such as communication, modelling and mathematical thinking (Kilpatrick, 2014; Niss, Bruder, Planas, Turner, & Villa-Ochoa, 2016). Kilpatrick (2014) attempted to define competency frameworks within the field of mathematics education, describing such frameworks as “[a] structural plan for organizing the cognitive skills and abilities used in learning and doing mathematics” (p. 85). The many frameworks divide mathematical competence into a set of mental processes, whilst emphasising that mastering mathematics requires a variety of competencies, not limited to carrying out procedures and memorising facts (Kilpatrick, 2014; Niss et al., 2016). The emergence of competency frameworks coincides with the increased focus on the acquisition of competencies in education (Pikkariainen, 2014; Westera, 2001), and competencies and competence-based curricula have been embraced “as a new standard for curriculum design” (Westera, 2001, p. 75). A parallel emergence of assessment frameworks can be observed, such as the frameworks developed for large international comparative studies (Kilpatrick, 2014; Turner, Dossey, Blum & Niss, 2013). One framework in particular has influenced mathematics curricula and assessment reforms in several European countries: the Danish KOM project (Kilpatrick, 2014). This framework comprises eight mathematical competencies that encapsulate what it means to master mathematics (Kilpatrick, 2014; Niss & Højgaard, 2002, 2011).<sup>1</sup>

When the aim of mathematics education is mathematical competence, teaching activities must provide students with opportunities to develop such competence and, in the words of Niss and Højgaard (2011), must be orchestrated “with the explicit aim of developing the mathematical competencies of the individual” (p. 31). However, the implementation of competency frameworks in national curricula does not necessarily have the intended effect on teaching and learning (Boesen et al., 2014; Charalambous & Philippou, 2010). Boesen et al. (2014) observed 197 Swedish classrooms to

<sup>1</sup> Niss & Højgaard (2011) is the English translation of Niss & Højgaard Jensen (2002). From here on, we will refer to Niss & Højgaard (2011).

investigate the impact of the introduction of competency goals in the Swedish national curriculum. They found that 15 years after the implementation of the new curricula, “classroom practice is still dominated by carrying out procedures” (Boesen et al., p. 85), occupying nearly 80% of the observed classroom time. Furthermore, the development of procedural competency was most dominant in classroom situations in which students worked individually or in groups with solving tasks. Although the research of Boesen et al. (2014) is situated in Sweden, a similar narrow focus might exist in other educational systems.

## Identifying What Matters: Analysing Tasks to Identify Competency Demand

In mathematics education, mathematical tasks play a prominent role (Kilpatrick, Swafford & Findell, 2001; Krauss, Baumert & Blum, 2008; National Council of Teachers of Mathematics [NCTM], 2000; Shimizu et al., 2010) because many teaching and learning activities in the mathematics classroom are based on solving mathematical tasks (Boesen et al., 2014; Doyle, 1988). Several studies show the importance of engaging students in cognitively demanding and complex tasks to promote higher learning outcomes (e.g. Boaler & Staples, 2008; Stein & Lane, 1996) and to develop students’ mathematical competence (Blomhøj & Jensen, 2007). Selecting appropriate tasks might be challenging for many teachers. Previous studies have shown that teachers tend to focus on surface features, such as mathematical content or topics, when analysing tasks (e.g. Arbaugh & Brown, 2005; Stein, Baxter & Leinhardt, 1990) and that teachers’ selection of instructional tasks is largely based on the concepts and skills that they need to cover (Hiebert et al., 1997).

Since 2003, the PISA Mathematics Expert Group (MEG) has been building on the mathematical competence framework developed in the KOM project (Niss, 2015) to develop and refine an item<sup>2</sup> analysis scheme. This scheme has been used to identify and rate the mathematical competencies needed to solve mathematical problems targeted at 15-year-old students (Turner, Blum & Niss, 2015) and those necessary to succeed in college courses (Tucker, 2013). Applying the scheme to analyse 48 mathematics items used in both the PISA 2003 and PISA 2006 surveys, Turner et al. (2013) found that it could be used effectively by experts to identify the competency demands of PISA items. Turner et al. (2015) also argue that the mathematical competencies described in the scheme “should legitimately be taking a prominent place in mathematics teaching and learning” (p. 108) and that the scheme can be used by teachers to select and devise appropriate assessment items. Indeed, for teaching and assessment to be aligned, teachers should be able to analyse different types of mathematical tasks to investigate whether the selected tasks offer opportunities to develop or demonstrate the competencies described in the curriculum. Other previous studies have demonstrated that tools and frameworks for analysing the cognitive demands of mathematical tasks can support teachers in thinking more deeply about mathematical tasks (e.g. the types of thinking

<sup>2</sup> An item refers to a question or task on a test that provides information about the test takers’ abilities or attributes. In this paper, the terms ‘item’ and ‘assessment task’ are used interchangeably.

that tasks promote; see, for instance, Arbaugh & Brown, 2005) and improve teachers' ability to select high-level tasks (Boston & Smith, 2011).

## Methodology

The aim of the present study was to investigate a group of five teachers' and prospective teachers' analysis of mathematical assessment tasks. After initial training on how to use the MEG item analysis scheme, they analysed the competency demands of 141 tasks from the 2014 Norwegian National Mathematics Grade 10 Exam ( $n_e = 56$ ) and the PISA 2012 survey ( $n_p = 85$ ). Our study focuses on the outcome of the teachers' and prospective teachers' analysis of the assessment tasks and how consistently they rated the six competencies whilst applying the item analysis scheme.

## Participants

The participants in this study were purposively recruited in January 2015 using a criterion sampling approach (Gall, Gall & Borg, 2007). The inclusion criteria were (1) some experience of teaching mathematics in secondary school and (2) having a degree in or being enrolled in a master's programme in secondary school mathematics teacher education.

Mathematics teachers who had previously been engaged in project work at the university were approached for recruitment. Colleagues with teacher experience and students enrolled in the master's programme in mathematics education at the university were also approached. These students had completed the teacher practice, consisting of 85 days of practice in lower or upper secondary schools, which is part of the master's programme, and thus met both inclusion criteria. In total, five teachers and prospective teachers (in the following sections referred to as 'teachers') participated in the study:

- One university employee with a PhD in mathematics education and prior experience (2 years) teaching in a secondary school
- One mathematics teacher at a secondary school with 18 months of full-time teaching experience, in addition to previous part-time teaching experience
- Three students enrolled in their final year of the mathematics education master's programme, of whom two worked as part-time mathematics teachers in upper secondary schools

## Material

The item analysis scheme consists of operational definitions of six mathematical competencies: Communication (C), Devising Strategies (DS), Mathematising (M), Representation (R), Using Symbols, Operations and Formal Language/Symbols and Formalism (SF), and Reasoning and Argument (RA) (the scheme can be found in full in Turner et al., 2015). In addition, four levels of demand (0 – 3) are described for each competency. When analysing a mathematical task, the competency level that best fit the demand of the task should be identified for each of the six competencies. A higher

rating indicates a higher cognitive demand; a competency rated at level 0 implies that the task does not demand activation of this competency (or only to a minimal degree), whilst level 3 implies an advanced or complex level of demand.

Assessment tasks from two mathematical tests were applied for the study: (1) 56 items from the 2014 Norwegian Grade 10 National Exam, consisting of part 1 (33 items), mainly comprising traditional tasks focused on applying procedures, and part 2 (23 items), emphasising problem-solving, and (2) 85 items from the paper-based PISA 2012 survey, aimed at assessing the degree to which students are able to apply mathematics to solve contextualised problems. Both the exam and PISA survey were targeted at 15-year-old students at the end of their compulsory education.

## Procedures

All teachers were proficient in English and were provided with the original English version of the MEG item analysis scheme in addition to a user guide that presents and explains the scheme. They were also given examples and explanations of item analysis performed by the MEG members (please see Turner et al., 2015, for this text). The five teachers were asked to spend 2 h familiarising themselves with the material prior to the 1-day training session that was focused on understanding and applying the item analysis scheme. The training consisted mainly of individually analysing PISA items from previous cycles, rating the demand of one competency at a time, followed by a group discussion of the analysis to promote a mutual understanding of the competency and level descriptions in the scheme. Following the training session, the teachers individually analysed all assessment items using the item analysis scheme. All completed their analysis within the following 3 weeks.

## Data Analysis Procedures

Several approaches were used to investigate the outcome of the teachers' analysis of the competency demands of the assessment tasks. The consistency in the teachers' analysis was examined to investigate the extent to which the five teachers identified the same competency demands for the same tasks. This was analysed by estimating both the internal consistency reliability (subsequently referred to as 'reliability') and interrater agreement as reliability and agreement estimates provide different information about the set of ratings (Tinsley & Weiss, 2000). Reliability was estimated through the Cronbach's alpha coefficient (Cronbach, 1951, p. 331), whilst agreement was estimated through an intraclass correlation coefficient (ICC) applying a two-way analysis of variance (Shrout & Fleiss, 1979). Fleiss and Cohen (1973) showed that ICC may be viewed as a special case of weighted kappa. Consequently, the discussions of the ICC estimates followed the guidelines proposed by Landis and Koch (1977), in which 0.21 – 0.40 represented fair agreement, 0.41 – 0.60 moderate agreement, 0.61 – 0.80 substantial agreement and 0.81 – 1.00 an almost perfect agreement. Whilst the interrater agreement indicates the degree to which the teachers give exactly the same ratings, the reliability is calculated using correlational indices and is therefore sensitive only to the ordering of the items (Tinsley & Weiss, 2000). For instance, if one teacher rates the demand for the communication competency for three items at levels 2, 0 and 1 and a second teacher rates the same items at levels 3, 1 and 2, the reliability between the two

sets of ratings would be a perfect 1, whilst the agreement measure would be less than 1. Clearly, applying both measures provides more information about the teachers' analysis.

Next, the distribution of ratings was calculated to provide information about the degree to which the teachers employed all four levels in their item analysis, to examine whether they were able to differentiate between different levels of competency demand. This may provide evidence regarding the five teachers' ability to apply the item analysis scheme to discriminate between high- and low-demand tasks.

## Results

The five teachers analysed all mathematics items used for the 2014 Norwegian national exam and the paper-based PISA 2012 survey. Figure 1 displays one of the released items<sup>3</sup> from the PISA 2012 survey, and Table 1 presents the outcome of the five teachers' analysis of the same item.

The ratings presented in Table 1 provide some insights into the teachers' utilisation of the item analysis tool, such as how they interpret the descriptions provided for the Communication competency. The level 2 description for this competency reads:

Identify and select elements to be linked, where repeated cycling within the material presented is needed to understand the task; or understand multiple elements of the context or task or their links. Any constructive communication involves providing a brief description or explanation, or presenting a sequence of calculation steps. (Turner et al., 2015, p. 111)

Before they can solve the question in 'Sailing Ships' (Fig. 1), students need to identify and link the relevant information from different elements in the item text. This is covered in the first sentence in the level description. Furthermore, students must support their answer by showing the calculations step by step, which fits well with the last sentence in the level description. For this item, the level 2 description for the Communication competency fits well with the item format and context, making the analysis relatively straightforward. For other items, the corresponding analysis might be more challenging because more in-depth analysis might be necessary to identify the appropriate competency level. In such cases, it is more likely that differences in the teachers' ratings will occur.

As can be observed from Table 1, the teachers seem to be fairly consistent in their analysis of 'Sailing Ships'. All teachers rate several of the competencies at level 2, indicating that they judge this item to be rather demanding. According to test results, this seems to be the case; across Organization for Economic Co-operation and Development (OECD) countries, 15% of the students successfully solved 'Sailing Ships', demonstrating that this was a fairly challenging item for 15-year-old students. Furthermore, Table 1 shows that the teachers agree that the item fits the level 0 description for Representation, indicating that they all consider that this item does not require activation of this competency. In addition, the teachers agree that the item fits the level 2 description for both Symbols and Formalism and Communication.

<sup>3</sup> <http://www.oecd.org/pisa/pisaproducts/pisa2012-2006-rel-items-maths-ENG.pdf>.




## Question 4: SAILING SHIPS

PM923Q04 – 0 1 9

Due to high diesel fuel costs of 0.42 zeds per litre, the owners of the ship *NewWave* are thinking about equipping their ship with a kite sail.

It is estimated that a kite sail like this has the potential to reduce the diesel consumption by about 20% overall.

|   |   |
|---|---|
| Name: <i>NewWave</i>  |  |
| Type: freighter   |   |
| Length: 117 metres  |   |
| Breadth: 18 metres  |   |
| Load capacity: 12 000 tons  |   |
| Maximum speed: 19 knots   |   |
| Diesel consumption per year without a kite sail: approximately 3 500 000 litres |   |

The cost of equipping the *NewWave* with a kite sail is 2 500 000 zeds.

After about how many years would the diesel fuel savings cover the cost of the kite sail? Give calculations to support your answer.

**Fig. 1** PISA 2012 mathematics assessment item ‘Sailing Ships’

As might be expected, the teachers did not always agree on the level of competency demand; examples of their varying analyses are also found in Table 1. Whilst four of the five teachers rate the Devising Strategies competency at level 2, teacher A regards this item as more demanding and assigned it a level 3. A similar pattern can be observed for the Mathematising competency, where teacher A again rates the level slightly higher than the other teachers do. Teacher C seems to judge the item to be less demanding than the other teachers do, rating the Reasoning and Argument competency at level 1, one level below the others. Nonetheless, the overall picture is that, although some differences are observed, the teachers agree on the question being demanding for several competencies.

**Table 1** Teachers’ ratings of the competency demands of the PISA item ‘Sailing Ships’ shown in Fig. 1

| Competency             | Teacher A | Teacher B | Teacher C | Teacher D | Teacher E |
|------------------------|-----------|-----------|-----------|-----------|-----------|
| Communication          | 2         | 2         | 2         | 2         | 2         |
| Devising Strategies    | <b>3</b>  | 2         | 2         | 2         | 2         |
| Mathematising          | <b>2</b>  | 1         | 1         | 1         | 1         |
| Representation         | 0         | 0         | 0         | 0         | 0         |
| Symbols and Formalism  | 2         | 2         | 2         | 2         | 2         |
| Reasoning and Argument | 2         | 2         | <b>1</b>  | 2         | 2         |

Deviated ratings are in bold



## Recognising the Competencies at Play: Consistency in the Teachers' Analysis

Ideally, teachers would analyse the competency demands with perfect agreement, identifying the 'true' competency demands of each item. However, in reality, as exemplified by Table 1, this cannot be the case. Previous studies have shown that analysing task demand may be challenging for teachers (Arbaugh & Brown, 2005; Stein et al., 1990). Moreover, teachers' subjective judgement may be biased by other factors; for instance, the so-called halo effect (Borman, 1975; Hoyt & Kerns, 1999) could result in items perceived as difficult being considered as highly demanding in all categories. In this study, this would mean that the teachers rate all or most competencies at level 2 or 3. This effect may increase if the categories in the item analysis scheme are viewed as abstract or unclear (Feeley, 2002).

Table 2 displays the Cronbach's  $\alpha$  coefficients of the teacher ratings for each competency, calculated for all items in total and for the exam and PISA items separately. An  $\alpha$  value of 1 implies a perfect correlation between the teachers' ratings of the competency demands of the items, whilst 0 would imply no correlation between the teachers' ratings.

Looking across all items, Table 2 indicates a fairly high reliability, with  $\alpha$  values greater than .80 for all six competencies, values ranging from .81 (Mathematising) to .89 (Devising Strategies). According to Cortina (1993), several studies consider  $\alpha > .70$  as acceptable without further interpretation. However, as  $\alpha$  is a function of many variables, e.g. the number of items in the scale, such use of absolute cutoff values for accepting  $\alpha$  without interpretations "is clearly an improper usage of the statistic" (Cortina, 1993, p. 101). In this study, the five teachers constitute the scale measuring the competency demand of each item. The values of  $\alpha > .80$  for a five-item scale (i.e. the five teachers) indicate a rather high reliability, that is, a high correlation between the teachers' ratings. When comparing  $\alpha$  values for the exam and PISA items separately, lower values are observed for the PISA items for all but the Representation competency. The lowest reliability is observed for the Reasoning and Argument competency for the PISA items, with an  $\alpha$  of .75. However, an  $\alpha$  of .75 still indicates a fairly high reliability. By estimating the reliability across all six competencies, we find a significantly higher value for the exam items ( $\alpha = .89$ , 95% CI = 88 – 91) than for the PISA items ( $\alpha = .84$ , 95% CI = 82 – 86).

The reliability estimates in Table 2 show that the teachers had similar opinions about which items should be rated high or low regarding the competency demands. To examine the absolute agreement of the teachers' ratings, ICCs were estimated by applying a two-way analysis of variance (Shrout & Fleiss, 1979). When estimating

**Table 2** Cronbach's  $\alpha$  coefficients from the item analysis of the assessment items

|            | C   | DS  | M   | R   | SF  | RA  |
|------------|-----|-----|-----|-----|-----|-----|
| All items  | .86 | .90 | .81 | .86 | .85 | .85 |
| Exam items | .90 | .92 | .85 | .81 | .84 | .89 |
| PISA items | .79 | .86 | .78 | .86 | .83 | .75 |

C Communication, DS Devising Strategies, M Mathematising, R Representation, SF Symbols and Formalism, RA Reasoning and Argument

ICCs, both average and single measurements can be calculated (McGraw & Wong, 1996). The average measure indicates the trustworthiness of the average ratings of the teachers seen as a group. At the same time, each teacher on his or her own might be less able to identify the competency demand of mathematical tasks. The single measure provides information about what the situation would look like if only one teacher analysed each task, indicating the extent to which we might rely on the analysis of a single teacher to represent the true competency demands of the item. Table 3 displays the single and average ICC measures for each competency for all items and for the exam and PISA items separately.

The average measures of agreement across assessments displayed in Table 3 range from .80 (Mathematising) to .88 (Devising Strategies), which means a substantial to an almost perfect agreement that strongly resembles the reliability estimates in Table 2. This indicates that the teachers were rather consistent in their ratings of competency demands across all 141 assessment tasks and, as such, have a fairly mutual understanding of the descriptions provided in the item analysis scheme. In addition, the average ratings of these five teachers should provide reliable information about the competency demands of the tasks, provided that the teachers operate as a group.

The lowest single measure is observed for the Mathematising competency, with an ICC of .44, which can be regarded as moderate agreement, according to Landis and Koch (1977). For the five other competencies, there is moderate to substantial agreement, ranging from .51 to .61. The ICC estimates for the single measures are lower than for the average measures. This indicates that in some teaching–learning situations, such as when planning a high-stakes assessment, the analysis made by an individual teacher might not be sufficiently reliable and accurate. Instead, in order to devise a test aimed at assessing students at different competency levels, engaging a group of teachers to analyse the tasks could be more constructive. The analysis of an individual teacher could be the starting point of a fruitful discussion amongst teacher colleagues, not only for assessment development but also for planning teaching activities.

Tables 2 and 3 show that for most of the competencies, the teachers' analysis was more consistent for the exam items than for the PISA items. For the exam items, the average agreement across all competencies was .89 (95% CI = .87 – .91), whilst for the PISA items the corresponding agreement was significantly lower, at .84 (95% CI = .82 – .86). Comparing the two parts of the exam, part 1 shows a higher agreement than part 2 for all competencies except Reasoning and Argument. By estimating the

**Table 3** Agreement measures (ICCs) of the teachers' competency ratings ('single teacher' measures are given in parentheses)

|             | C         | DS        | M         | R         | SF        | RA        |
|-------------|-----------|-----------|-----------|-----------|-----------|-----------|
| All items   | .86 (.55) | .88 (.61) | .80 (.44) | .84 (.51) | .85 (.52) | .84 (.51) |
| Exam items  | .89 (.61) | .92 (.69) | .84 (.50) | .79 (.43) | .82 (.47) | .89 (.61) |
| Exam part 1 | .86 (.54) | .92 (.69) | .89 (.62) | .86 (.54) | .85 (.52) | .78 (.42) |
| Exam part 2 | .79 (.43) | .88 (.60) | .71 (.34) | .67 (.29) | .79 (.42) | .84 (.52) |
| PISA items  | .77 (.40) | .86 (.54) | .77 (.39) | .84 (.50) | .83 (.49) | .74 (.36) |

*C* Communication, *DS* Devising Strategies, *M* Mathematising, *R* Representation, *SF* Symbols and Formalism, *RA* Reasoning and Argument

agreement measures across all six competencies, the average ICC is .90 (95% CI = 88 – 92) for part 1 and .84 (95% CI = 80 – 88) for part 2. However, these measures are not significantly different at the .05 level, partly due to the relatively small number of items in part 1 ( $n = 33$ ) and part 2 ( $n = 23$ ).

The two assessments aim at assessing different types of mathematical knowledge and skills. The PISA 2012 survey aims at assessing mathematical literacy, including a modified formulation of the Niss competencies, as defined in the PISA framework (OECD, 2013a). This is visible in the item in Fig. 1 that, according to the teachers' analysis, demands the use of several mathematical competencies. The main goal of the national exam is to assess students' mathematical competence as defined by the national curricula. However, the two parts of the exam have different profiles and assess different aspects of mathematical competence. Part 1 mainly aims at assessing students' procedural skills and automated knowledge, and part 2, to a large extent, aims at assessing both the width and depth of students' mathematical competencies (Norwegian Directorate for Education and Training, 2014), comprising items that seem to be more similar to the PISA assessment. The PISA item in Fig. 1 can serve as an example of an item that would also have fitted the profile of part 2 of the exam.

The fraction item shown in Fig. 2 represents a typical part 1 exam item. The fraction item comprises a short sentence explaining what to do and an arithmetic task to be solved. Table 4 displays the analysis of the fraction item, showing that the teachers agree that this item draws at a single competency (Symbols and Formalism) and at a rather low level.

### Challenges in Applying the Item Analysis Tool

Whilst some differences in teacher ratings were observed for the item 'Sailing Ships', the five teachers rated the fraction item identically. A striking difference between the fraction item and 'Sailing Ships' is the number of activated competencies. In the context of this study, 'Sailing Ships' might be termed a complex task. A complex task is a task that demands activation of multiple competencies. One hypothesis is that tasks that are more complex are more difficult to analyse and may contribute to less consistent ratings or halo effects. The complexity of items in the exam and the PISA test was estimated calculating the average number of competencies rated above level 0 for the items comprised by each assessment. These estimations show that, on average, a higher number of competencies were rated above level 0 for the PISA items (3.97 competencies/item) than for the exam items (2.96 competencies/item). Thus, on average, the PISA items can be regarded as more complex. However, as part 1 and part 2 of the exam serve different purposes, separate estimates for the two parts could give a more accurate image of the complexity of the exam. Indeed, the difference in

Regn ut, og forkort brøken hvis det er mulig

$$\frac{5}{2} - \frac{2}{3} = \underline{\hspace{2cm}}$$

**Fig. 2** Fraction item from exam part 1: "Calculate and simplify the fraction if possible". (Authors' translation)

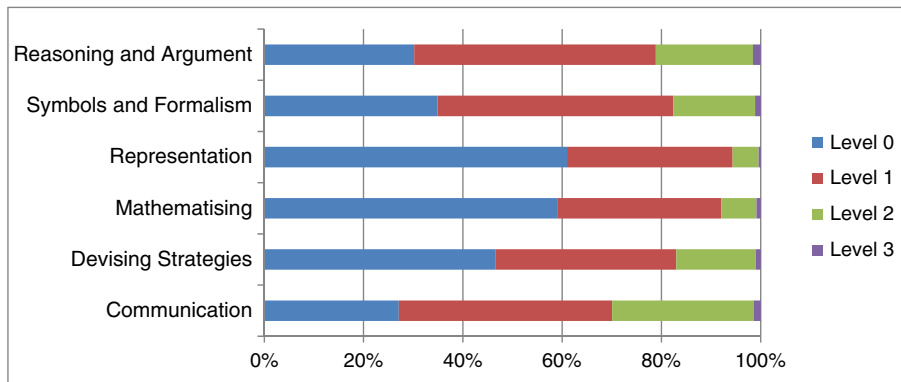
**Table 4** Teachers' ratings of the competency demands of the fraction item in Fig. 2

| Competency             | Teacher A | Teacher B | Teacher C | Teacher D | Teacher E |
|------------------------|-----------|-----------|-----------|-----------|-----------|
| Communication          | 0         | 0         | 0         | 0         | 0         |
| Devising Strategies    | 0         | 0         | 0         | 0         | 0         |
| Mathematising          | 0         | 0         | 0         | 0         | 0         |
| Representation         | 0         | 0         | 0         | 0         | 0         |
| Symbols and Formalism  | 1         | 1         | 1         | 1         | 1         |
| Reasoning and Argument | 0         | 0         | 0         | 0         | 0         |

complexity is even more pronounced for the two parts of the exam. Items in part 1, the more traditional tasks emphasising procedural knowledge, mainly comprising the Symbols and Formalism competency, on average demand 2.15 competencies/item, whereas in part 2, which aims at measuring both the breadth and depth of students' mathematical competence, the average number of competencies demanded per item is almost twice as many (4.13). The observed differences in item complexity correspond to the differences in reliability and agreement shown in Tables 2 and 3: the more complex items in PISA and exam part 2 show a lower consistency in the teachers' analysis than do the less complex items in exam part 1. Thus, these results support the hypothesis that the more complex items were more challenging to analyse for the five teachers, which led to less consistency in their ratings.

The overall aim of this study was to investigate the outcome of the teachers' analysis of assessment tasks when utilising an item analysis scheme. The previous sections demonstrate that the teachers agreed, to a large extent, on which competencies students need to activate in order to solve the assessment items. When planning their teaching activities, teachers must be able to recognise competency demands at different levels to select appropriate tasks for students who need different challenges to support their learning. As previous studies have shown that cognitively demanding tasks promote higher learning outcomes (Boaler & Staples, 2008; Stein & Lane, 1996), it is important that teachers are able to differentiate between low and high levels of demand. In the present study, each competency was rated 141 times by each of the five teachers. Consequently, the distribution of the ratings across the four levels could give some indication of the five teachers' ability to identify both low and high competency demand whilst applying the item analysis scheme.

Figure 3 displays the distribution of the teachers' ratings for each of the six competencies; the majority of the ratings are at levels 0 and 1, indicating that the teachers judge the majority of items to comprise a low level of cognitive demand. In addition, Fig. 3 indicates that the five teachers rarely identify tasks as highly demanding; level 3 is hardly used. In addition, less than 5% of the Mathematising and Representation ratings are at level 2. Two plausible explanations for the observed pattern might be found: (1) the five teachers struggle to recognise high levels of competency demand when applying the scheme to analyse assessment items or (2) the two assessments comprise few highly demanding items. Regarding the latter explanation, some external evidence exists that confirm that both assessments comprise



**Fig. 3** Distribution of the total number of ratings for all items, on each of the four levels for the six competencies (705 ratings/competency)

highly demanding items. For instance, the PISA 2012 mathematics assessment encompasses items at six proficiency levels, in which items connected to the two highest levels assess complex problem-solving (OECD, 2013b). All PISA 2012 mathematics items were included in the present study. Furthermore, 48 of the included items were previously analysed in the MEG Study; when comparing the teachers' analysis to those of the PISA MEG, the reports from the MEG Study (Turner et al., 2013) indicate a higher proportion of ratings at levels 2 and 3.

Thus, the observed pattern in Fig. 3 is indicative of the teachers' difficulties in recognising the high-level competency demands described in the item analysis scheme. This could indicate that the level descriptions are inadequate operationalisations of the actual competency demands of mathematics items. Consequently, the teachers may have struggled to understand and differentiate between the higher levels. For instance, relative terms, such as 'simple' and 'complex', are used in the level descriptions to define the difference between activation at a low and high level. These words tend to have different meanings for different people (Turner et al., 2015). During the revisions of the item analysis scheme, Turner et al. (2015) attempted to minimise the use of such terms. Nonetheless, when examining the wording of levels 2 and 3, we find that the term 'complex' is frequently used in the descriptions, as shown in the following excerpt from the description of level 3 for the Representation competency: "Understand, use, link or translate between multiple complex representations of mathematical entities; or compare or evaluate representations; or devise a representation that captures a complex mathematical entity" (Turner et al., 2015, p. 113). In addition, the duration of the training period in the present study may have been too short to provide the teachers with the in-depth understanding required to differentiate between the four levels in the rating scale. A longer and more exhaustive training session might have strengthened the teachers' understanding of the competencies and level description. The five teachers all had rather limited teaching experience. More experience might contribute to teachers' ability to recognise what a complex competency demand is for 15-year-old students, partly because teachers with more experience have had more opportunities to observe how students at different ability levels engage with mathematical tasks.

## Discussion and Concluding Remarks

According to Niss and Højgaard (2011), students develop mathematical competence through the activities and tasks in which they engage. The prominent role of mathematical tasks in the teaching and learning of mathematics is well established in the research literature (see, for instance, Hiebert & Wearne, 1993; Kilpatrick et al., 2001; NCTM, 2000; Shimizu et al., 2010; Stein, Smith, Henningsen & Silver, 2000). Consequently, teachers must be able to recognise and select appropriate tasks to be used in their teaching to stimulate their students' development of mathematical competence (Kilpatrick et al., 2001; Niss & Højgaard, 2011; Shimizu et al., 2010; Turner et al., 2015). Previous studies indicate that whilst teachers tend to focus on surface features of mathematical tasks (Arbaugh & Brown, 2005; Stein et al., 1990), training may develop teachers' abilities to analyse the cognitive demand of tasks (Arbaugh & Brown, 2005; Boston & Smith, 2011) and hence enable them to improve their students' opportunities for learning.

The present study sought to explore the outcome of five teachers' and prospective teachers' analysis of assessment tasks when applying an item analysis scheme. The main finding is that the teachers could apply the scheme to identify which competencies 15-year-old students need to activate to solve the item, but not to differentiate between high and low competency demand. Some limitations to the present study need to be addressed before further discussing the results. First, only assessment items were included in the study. This is an important restriction because assessment items represent only one group of the wide variety of tasks teachers should utilise for teaching and learning situations. The main rationale for selecting exam and PISA items is that these are developed to assess the mathematical competence of students at the end of compulsory education. Both 'traditional' and more open-ended and complex items are represented in the sample, and as such, they represent different kinds of tasks with which teachers should be familiar. When planning for teaching sessions, teachers will encounter novel tasks, and the selected material represents a realistic situation in which teachers have to analyse tasks that are previously unknown to them. Next, the restricted number of participants, and the fact that all participants are connected to the same university department, means that the results are not, per se, generalisable to other teachers and prospective teachers. That is, other teachers and prospective teachers might have taken more or less training to familiarise themselves with the item analysis scheme, or they might have agreed to a greater or lesser degree on the competency demands of the provided items. Finally, the item analysis scheme was only provided in English. Seemingly, this was not an issue during the training session; the participants' challenges with the scheme were connected to the use of terms such as 'complex'. These terms are also used in the same manner in the Norwegian language (*kompleks*) and would have been equally unclear in a Norwegian translation of the scheme.

Regarding the research question, the outcome of the present study demonstrates that the teachers as a group analysed the competency demands of the assessment tasks rather consistently. This indicates that they have a fairly mutual understanding of the competencies and the extent to which the different competencies need to be activated to solve the tasks successfully. However, the results also indicate that differentiating between four levels of competency demand proved challenging. This may be partly due to the training session, which may have been too short for the participants to

develop a deep understanding of the mathematical competencies and competency levels. Such a deep understanding may take considerable time and exposure to develop. In addition, the teachers may have struggled with recognising what complex competence looks like in a 15-year-old student. This is connected to the teachers' proficiency in teaching mathematics. These explanations are supported by the outcome of the MEG Study, where the analysis shows more use of the full rating scale (Turner et al., 2013). The PISA MEG comprised experienced members who had worked together for many PISA cycles and, as such, had time to refine their understanding of the scheme. In contrast, the teachers and prospective teachers in the present study can be regarded as novices, both in relation to the item analysis scheme and to teaching. It is likely that, with more training and experience, the teachers and prospective teachers would have been able to further refine their understanding of the scheme and of what characterises complex competency demand in tasks aimed at 15-year-old students. Still, the ability to recognise different levels of demand in mathematical tasks is important for novice teachers, too, and as such should be stressed during teacher training.

In addition, the more complex assessment tasks were seemingly more challenging to analyse for the five teachers and prospective teachers. This inference is supported by the low occurrence of ratings at levels 2 and 3, as well as by the differences in agreement between, on the one hand, exam part 1 and, on the other, the more complex items in exam part 2 and PISA items. Previous research has revealed that tasks in mathematics classrooms focus mainly on applying procedures (Boesen et al., 2014; Hiebert et al., 2003; Kaur, 2010; Palm et al., 2011). This could indicate that the teachers and prospective teachers are less familiar with tasks that are more complex and require the activation of multiple competencies. This may be a concern since cognitively demanding tasks seemingly promote higher learning outcomes (Boaler & Staples, 2008; Stein & Lane, 1996) and thus should play a considerable role in mathematics education.

Whilst the teachers and prospective teachers as a group analysed the provided tasks consistently, the agreement measures also revealed that the analysis of an individual teacher was less accurate and reliable. Thus, as a collaborative activity amongst mathematics teachers or prospective teachers, the scheme could be a valid tool for promoting and enhancing discussions and reflections about the competency demands of mathematical tasks. As such, the item analysis scheme could be used in teacher training or for professional development through teacher collaborations within schools. This type of professional development has been proven fruitful in previous studies (see, for instance, Arbaugh & Brown, 2005; Boston & Smith, 2011), where critical examination of mathematical tasks supports growth in pedagogical content knowledge and changes teachers' practice. Further revision of the item analysis scheme may be useful to help teachers better understand the level descriptions, to support them in differentiating between high and low competency demands in mathematical tasks.

**Acknowledgements** The authors would like to thank Mogens Niss and Ross Turner from the PISA MEG for their support and discussions prior to the data collection for this study. Further, the teachers and prospective teachers are thanked 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. A short version of this article has been accepted for presentation at the 40th Annual Meeting of the International Group for Psychology of Mathematics Education (PME 40), Szeged, Hungary (Pettersen & Nortvedt, 2016).



## References

- Arbaugh, F. & Brown, C. A. (2005). Analyzing mathematical tasks: A catalyst for change? *Journal of Mathematics Teacher Education*, 8(6), 499–536.
- Blömeke, S., Gustafsson, J.-E. & Shavelson, R. J. (2015). Beyond dichotomies. *Zeitschrift für Psychologie*, 223(1), 3–13.
- Blomhøj, M. & Jensen, T. H. (2007). What's all the fuss about competencies? In W. Blum, P. L. Galbraith, H.-W. Henn & M. Niss (Eds.), *Modelling and applications in mathematics education: The 14th ICMI Study* (pp. 45–56). Dordrecht: Springer.
- Boaler, J. & Staples, M. (2008). Creating mathematical futures through an equitable teaching approach: The case of Railside School. *The Teachers College Record*, 110(3), 608–645.
- 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.
- Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology*, 60(5), 556–560.
- Boston, M. D. & Smith, M. S. (2011). A 'task-centric approach' to professional development: Enhancing and sustaining mathematics teachers' ability to implement cognitively challenging mathematical tasks. *ZDM*, 43(6–7), 965–977.
- Burkhardt, H. (2014). Curriculum design and curriculum change. In Y. Li & G. Lapan (Eds.), *Mathematics curriculum in school education* (pp. 13–33). Dordrecht: Springer.
- Charalambous, C. Y. & Philippou, G. N. (2010). Teachers' concerns and efficacy beliefs about implementing a mathematics curriculum reform: Integrating two lines of inquiry. *Educational Studies in Mathematics*, 75(1), 1–21.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1), 98–104.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.
- Doyle, W. (1988). Work in mathematics classes: The context of students' thinking during instruction. *Educational Psychologist*, 23(2), 167–180.
- Feeley, T. H. (2002). Comment on halo effects in rating and evaluation research. *Human Communication Research*, 28(4), 578–586. doi:10.1111/j.1468-2958.2002.tb00825.x.
- Fleiss, J. L. & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3), 613–619.
- Gall, M. D., Gall, J. P. & Borg, W. R. (2007). *Educational research: An introduction* (8th ed.). Boston: Allyn and Bacon.
- Hiebert, J. (1986). *Conceptual and procedural knowledge: The case of mathematics*. Hillsdale: Erlbaum.
- Hiebert, J. & Wearne, D. (1993). Instructional tasks, classroom discourse, and students' learning in second-grade arithmetic. *American Educational Research Journal*, 30(2), 393–425.
- Hiebert, J., Carpenter, T. P., Fennema, E., Fuson, K., Wearne, D., Murray, H., Olivier, A. & Human, P. (1997). *Making sense: Teaching and learning mathematics with understanding*. Portsmouth, NH: Heinemann.
- Hiebert, J., Gallimore, R., Garnier, H., Givvin, K., Hollingsworth, H., Jacobs, J., Chui, A. M., Wearne, D., Smith, M., Kersting, N., Manaster, A., Tseng, E. A., Etterbeek, W., Manaster, C. & Stigler, J. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 video study*. NCES 2003–013. Washington: National Center for Education Statistics.
- Hoyt, W. T. & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, 4(4), 403–424.
- Kaur, B. (2010). Mathematical tasks from Singapore classrooms. In Y. Shimizu, B. Kaur, R. Huang & D. Clarke (Eds.), *Mathematical tasks in classrooms around the world* (pp. 15–33). Rotterdam: Sense Publishers.
- Kilpatrick, J. (2014). Competency frameworks in mathematics education. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 85–87). Dordrecht: Springer.
- Kilpatrick, J., Swafford, J. & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academies Press.
- Koepfen, K., Hartig, J., Klieme, E. & Leutner, D. (2008). Current issues in competence modeling and assessment. *Zeitschrift für Psychologie*, 216(2), 61–73.
- Krauss, S., Baumert, J. & Blum, W. (2008). Secondary mathematics teachers' pedagogical content knowledge and content knowledge: Validation of the COACTIV constructs. *The International Journal on Mathematics Education*, 40(5), 873–892. doi:10.1007/s11858-008-0141-9.

- Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.
- Lithner, J. (2004). Mathematical reasoning in calculus textbook exercises. *The Journal of Mathematical Behavior*, 23(4), 405–427.
- McGraw, K. O. & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30–46.
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston: National Council of Teachers of Mathematics.
- Niss, M. (2015). Mathematical competencies and PISA. In K. Stacey & R. Turner (Eds.), *Assessing mathematical literacy* (pp. 35–55). Dordrecht: Springer.
- Niss, M. & Højgaard, T. (Eds.). (2011). *Competencies and mathematical learning*. Roskilde: Roskilde University.
- Niss, M. & Højgaard J. T. (2002). *Kompetencer og matematiklæring: ideer og inspiration til udvikling af matematikundervisning i Danmark* [Competencies and mathematical learning: ideas and inspiration for the development of mathematics teaching and learning in Denmark] (Vol. nr 18). Copenhagen: Undervisningsministeriet.
- 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.
- Norwegian Directorate for Education and Training (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 (2013a). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris: OECD Publishing.
- Organization for Economic Co-operation and Development (2013b). *PISA 2012 result: What students know and can do—Student performance in Mathematics, Reading, Science (volume I)*. Paris: OECD Publishing.
- Palm, T., Boesen, J. & Lithner, J. (2011). Mathematical reasoning requirements in Swedish upper secondary level assessments. *Mathematical Thinking and Learning*, 13(3), 221–246.
- Pikkarainen, E. (2014). Competence as a key concept of educational theory: A semiotic point of view. *Journal of Philosophy of Education*, 48(4), 621–636.
- Shimizu, Y., Kaur, B., Huang, R. & Clarke, D. (2010). The role of mathematical tasks in different cultures. In Y. Shimizu, B. Kaur, R. Huang & D. Clarke (Eds.), *Mathematical tasks in classrooms around the world* (pp. 1–14). Rotterdam: Sense Publishers.
- Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428.
- Stein, M. K. & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation*, 2(1), 50–80. doi:10.1080/1380361960020103.
- Stein, M. K., Baxter, J. A. & Leinhardt, G. (1990). Subject-matter knowledge and elementary instruction: A case from functions and graphing. *American Educational Research Journal*, 27(4), 639–663.
- Stein, M. K., Smith, M. S., Henningsen, M. A. & Silver, E. A. (2000). *Implementing standards-based mathematics instruction: A casebook for professional development*. New York: Teachers College Press National Council of Teachers of Mathematics.
- Tinsley, H. E. A. & Weiss, D. J. (2000). Interrater reliability and agreement. In H. E. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95–124). San Diego: Academic.
- Tucker, M. (2013). *What does it really mean to be college and work ready? The mathematics required of first year community college students*. Washington, DC: National Center on Education and the Economy.
- 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.
- 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.
- Westera, W. (2001). Competences in education: A confusion of tongues. *Journal of Curriculum Studies*, 33(1), 75–88.