Methodological issues when studying the relationship between reading and solving mathematical tasks

MAGNUS ÖSTERHOLM AND EWA BERGQVIST

In this paper we examine four statistical methods used for characterizing mathematical test items regarding their demands of reading ability. These methods rely on data of students' performance on test items regarding mathematics and reading and include the use of regression analysis, principal component analysis, and different uses of correlation coefficients. Our investigation of these methods focuses on aspects of validity and reliability, using data from PISA 2003 and 2006. The results show that the method using principal component analysis has the best properties when taking into account aspects of both validity and reliability.

The relation between reading ability and mathematical ability is complex and important to examine. One obvious reason is that the predominant form of assessment in mathematics teaching is pen-and-paper tests. The ability to read and write is thus necessary in order to answer the test items correctly, or even to answer the items at all. It is often stated that such written tests should avoid measuring reading ability, in particular by using as simple language as possible (see e.g. OECD, 2003, 2006). At the same time it is a common view that communication is an important aspect of mathematical knowledge (e.g. NCTM, 2000; Niss & Jensen, 2002; OECD, 2003, 2006) and in particular that reading in mathematics demands a specific type of reading ability, which needs to be taught at all educational levels (Burton & Morgan, 2000; Cowen, 1991; Fuentes, 1998; Konior, 1993; Krygowska, 1969; Shanahan & Shanahan, 2008). Our conclusion is that it is not suitable, or even possible, to separate reading ability from mathematical ability and that further studies are needed about the intricate connections between them.

Magnus Österholm, *Umeå University and Monash University, Melbourne*
Ewa Bergqvist, *Umeå University*

Österholm, M. & Bergqvist, E. (2012). Methodological issues when studying the relationship between reading and solving mathematical tasks. *Nordic Studies in Mathematics Education*, 17 (1), 5–30.

The overarching goal of our research, not to be fulfilled in this particular paper, is to better understand and be able to describe the connections between reading and solving mathematical tasks. This goal includes many, and fundamentally different types of, sub-goals. One sub-goal is to develop a theoretical model of the process of reading and the process of solving mathematical tasks, and a first suggestion for such a theoretical model is presented by Bergqvist and Österholm (2010). Another sub-goal is to identify linguistic features of mathematical tasks that are particularly important when students solve these tasks, and we have also started to explore empirical data from this perspective (Bergqvist, Dyrvold & Österholm, in press; Österholm & Bergqvist, in press). Furthermore, one important aspect of the overarching goal is that we are not only interested in studying the relation between reading ability and mathematical ability. Instead, we focus mainly on the processes involved in reading and solving, although we also assume that there are predicting factors, such as reading ability and mathematical ability, which are related to the outcome of these processes. We therefore plan for many different types of studies using different methods and different types of data in order to gradually give us a deeper understanding of the connection between reading and solving mathematical tasks. Neither the conclusions concerning empirical methods presented in this paper nor results from our other studies are supposed to be used to analyze all aspects of task solving or all aspects of reading comprehension. Instead, the purpose is to capture the dynamic relationship between reading and solving tasks, in order to be able to understand this relationship in more detail.

There are many possible sources of data and methods of analysis that can be used in this quest. As a survey of previous research, we include a discussion of several different empirical methods for examining relationships between reading and solving mathematical tasks. Thereafter we focus our attention on a specific type of research method; the utilization of different sets of test results that include students taking both mathematics and reading tests of some kind. Such data can provide us with information about students' mathematical ability and reading ability, to be used as possible predictors of their outcome for each mathematical task. In the present paper we discuss four different statistical methods that are used for examining the connection between students' results on tests of mathematical ability and reading ability on the one hand and students' outcome on specific mathematical tasks on the other hand. We discuss the reliability and validity of these methods when applying them to data from PISA 2003 and 2006.

In conclusion, the main purpose of this paper is to discuss methodological issues in research about relationships between reading and solving mathematical tasks, and in particular to examine aspects of reliability and validity of methods intended to characterize mathematics test items regarding how dependent on reading ability students' outcomes are, in order to find out which method is to be preferred.

Background

Relationships between reading and solving mathematical tasks can be studied in many different ways, and here the discussion is organized around three areas of research; (1) the study of linguistic properties of mathematical tasks in relation to the solving of these tasks, (2) focusing on the process of reading and solving mathematical tasks, and (3) the use of quantified measures of different abilities (regarding reading and mathematics) in relation to the solving of mathematical tasks. The analysis in this paper focuses on the third area of research, primarily because of problems and limitations regarding the first two areas, as discussed below.

Research area 1: Effects of linguistic properties of mathematical tasks

The basic idea for studies within this first research area is to examine if and how different properties of a task affect the solving of the task. Of interest in this paper is when these properties are of linguistic type, thereby seen as connected to the reading of the task and (perhaps implicitly) primarily not seen as connected to the mathematical content of the task. Thus, studies in this area of research try to characterize mathematical tasks regarding the demands of reading ability by examining different aspects of linguistic complexity of tasks.

The literature survey by Österholm (2007, p. 142), focusing on studies about word problems, includes several empirical studies that show that "the performance in solving problems can be negatively affected by a higher complexity of the language used in the problem text". Complexity can involve different things, such as the number of difficult words or word length (see e.g. Homan, Hewitt & Linder, 1994) or the voice of verb phrases and the use of conditional clauses (see e.g. Abedi & Lord, 2001). However, there are also studies showing no effect of higher linguistic complexity on performance, for example regarding longer sentences (Muth, 1984).

Some researchers try to capture the degree of complexity by using readability formulas. However, a general problem is that such formulas are usually created for prose and longer texts than a test item, in particular that formulas usually require passages of at least 100 words (Homan et al., 1994). There have been attempts to create special kinds of formulas for test items. The values from some of these formulas show no connection to students' performance (Paul, Nibbelink & Hoover, 1986) while other formulas show a clear connection (Homan et al., 1994). When using several different kinds of formulas, Walker, Zhang, and Surber (2008, p. 177) conclude that ”the majority [of the formulas] produced such unreliable and inconsistent results that at times it seemed as if the numbers were obtained using a random number generator”. Also, based on a review of research about the use of readability formulas for test items, Oakland and Lane (2004, p. 250) draw the conclusion that ”the formulas should not be used at the item level until their reliability and validity are known”.

In addition to the problems described around the use of readability formulas, two other limitations can be seen for studies in this research area. A first limitation is that regarding some studies of linguistic properties of tasks it can be difficult to decide what is ”pure linguistics” and what is more connected to the mathematical aspects of the task. For example, in one study the task texts were altered in some word problems so that the relationship between known and unknown quantities became more explicit (Bernardo, 1999). The author describes this type of re-wording of tasks as a linguistic change, but this change could perhaps also be seen as affecting the validity of the task regarding its potential in testing a student's mathematical ability, for example a competence of mathematical communication or mathematical modeling. Another study takes into consideration that linguistic alterations also can change mathematical aspects of a test item and uses different methods to test if the mathematical properties of the task also change (Sato, Rabinowitz, Gallagher & Huang, 2010). Results from the different types of analyses are not always consistent, and the authors see a need for continued research that ”may lead to better understanding of the ways in which item and content characteristics interact with linguistic modification strategies” (Sato et al., 2010, p. 4).

A second limitation of studies in this research area is that they examine only properties of the task text, without considering the text in relation to a person who is reading and solving the task. For example, Oakland and Lane (2004) note that there is more to readability than language properties, for example text legibility and interest. In addition, there is research about reading comprehension of expository texts showing that it is not always better with ”simpler” texts. In particular, there are results showing that regarding the coherence of a text, it can be that ”readers who know little about the domain of the text benefit from a coherent text, whereas high-knowledge readers benefit from a minimally coherent text”

(McNamara, Kintsch, Songer & Kintsch, 1996, p. 1). Whether this effect also can be found for the reading of test items is a topic for future studies.

Research area 2:
The process of reading and solving mathematical tasks

Studies in the first research area, presented in the previous section, can be said to focus on input (text properties) and output (student performance). As a contrast, studies in the second area examine aspects of the processes of reading and solving mathematical tasks, for example regarding students' thinking, reasoning, or strategy use. There seems to exist few studies in this research area, for example as noted in Österholm's (2007) literature survey. However, some studies have examined aspects of the process of solving mathematical tasks from a reading perspective, and such studies are discussed here.

Based on data consisting of students' written solutions on mathematical tasks, Möllehed (2001) draws a conclusion that the most common type of error is caused by a lack of *textual understanding*, as he labels it. However, it seems difficult to draw such a conclusion based only on the students' written solutions, where a very indirect view of the thoughts and reasoning behind the written descriptions is given. For example, as a sign of lack of textual understanding, Möllehed includes both students' "arbitrary calculations" using the given numbers and also their way of solving a different, and perhaps simpler, task than the given task. However, these types of errors could also be explained by students' reliance on imitative or pseudo-analytical reasoning (see e.g. Lithner, 2008; Vinner, 1997). In addition, when Lager (2006) examined students' written solutions and also interviewed them, he noted that several students in the interview corrected their own prior errors and also that the interview gave insights in students' reasoning that were not evident from their written solutions.

Lager (2006), who focuses on second language learners, is an example of a researcher that has used interviews to, in a more direct manner, try to examine aspects of reading comprehension in relation to attempts to solve mathematics tasks. Knifong and Holtan (1977) also interviewed students about their comprehension of tasks that they had previously tried but failed to solve. The purpose of the interview was to examine if the students could read the text aloud and also describe the situation in the text and what was asked for in the given task. The results showed that the students almost all the time (in more than 90 % of the cases) had a good comprehension of the text but that they during the interview seldom (in 36 % of the cases) could describe a correct solution.

The empirical studies discussed so far have a similar focus; to examine how great impact (lack of) reading comprehension has on students' difficulties when solving mathematical tasks. Another type of study in this research area focuses on what type of solution strategies students use for different types of mathematical tasks (all arithmetical word problems), in particular if and how these strategies differ between students with different reading and mathematical ability (Søvik, Frostad & Heggberget, 1999). The authors' hypothesis was that differences in the use of solution strategies among students with different reading ability could be one part of explaining the common result showing a connection between reading ability and mathematical performance. However, the result showed that for all types of tasks, strategy use was independent of reading ability, while a difference in strategy use was noted regarding some types of tasks when comparing students with different mathematical ability. The difference was that students with higher mathematical ability tended to use a deductive strategy while students with lower mathematical ability tended to use a procedural strategy.

The type of studies in this second research area are not discussed more in this paper, mainly due to the lack of this type of studies and that the focus of existing studies are on more general aspects of student performance, and not on characterizing mathematical tasks. However, we see a great potential in developing studies in this research area as a complement to the other areas, for example in order to examine more thoroughly how different properties of tasks in conjunction with students' reasoning and strategy use can influence the reading and solving of the tasks – but we see a need for more methodological development before this can be realized.

Research area 3: Effects of different student abilities

Studies in this research area focus on input and output by examining student abilities as input and their performances on mathematical tasks as output. A main difference compared to the first research area is that focus is now on properties of students and not of tasks. However, a purpose with studies in this area is also to draw conclusions about properties of tasks, primarily by examining those tasks that show the highest demand of reading ability. Several different statistical methods have been used in different studies in order to examine how students' performance on mathematical tasks depends on the two abilities, regarding mathematics and reading. Before discussing different statistical methods, it is necessary to also discuss some theoretical aspects regarding different abilities, in particular relationships between them.

When discussing *abilities* we here refer to (relatively) stable types of traits of students, which are seen as latent variables that can be estimated through tests of different kinds. A significant correlation between mathematical ability and reading ability exists in many studies, for example among Swedish and Norwegian students in PISA 2003, with a correlation coefficient of 0.57 (Roe & Taube, 2006) and in plenty of older studies, with coefficients between 0.40 and 0.86 (Aiken, 1972). The literature survey by Aiken also shows that to a large extent, but not entirely, this correlation can be explained by a common reliance on a general cognitive ability (intelligence). These results highlight the difficulty to separate these two abilities. However, for the present study, a model is used that assumes the existence of two different latent variables, called mathematical ability and reading ability. This model is of course a simplification of reality, in particular regarding our awareness that the two abilities are not totally different (i.e. not separated) and also that each ability is not one-dimensional in itself (e.g. see OECD, 2006). However, our assumption when using this model is that the abilities are separated and homogenous enough in order to study them as two different dimensions of human cognition. It can of course be of interest to refine the discussion and analysis in this paper (and other studies) using a more fine-grained model of different competences within each of the two main abilities (as is done to some degree by Nortvedt, 2009), but this paper focuses on the two abilities more generally.

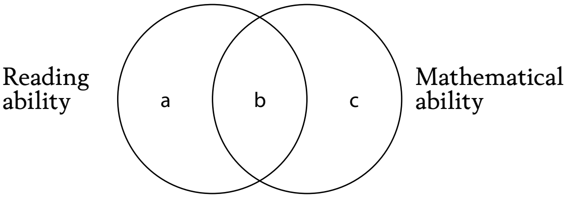

Figure 1. *Schematic characterization of relationships between abilities*

In conclusion, regarding theoretical considerations, in this paper we use a model consisting of two different abilities, mathematical and reading, with the following essential properties and relationships:

- Homogeneity: Each ability is homogenous enough to see it as *one* ability.
- Separation: The two abilities are separated enough to see them as two different abilities.
- Overlap: There is a considerable overlap between the abilities.

These properties are schematically depicted in figure 1, where our main interest in the present study is the effect of area (a); the genuine effect of reading ability on student performance on mathematics tasks, which in this study is denoted as a task's *demand of reading ability*.

The correlation between the result on a single mathematics task and reading ability is used by both Roe and Taube (2006) and Nortvedt (2009) in order to characterize tasks regarding how much they demand of the students' reading ability. With this method no consideration is taken of the overlap between mathematics ability and reading ability, which can be done using regression analyses. A regression analysis by Muth (1984), using computational ability and reading ability as explanatory variables for the results on a mathematics test with arithmetic word problems, shows that 7.6 % of the variance is explained uniquely by reading ability while 13.6 % is explained uniquely by computational ability and 32.4 % of the explained variance was common for the two abilities. In the analysis by Muth, (an aspect of) mathematical ability is included through the use of a regression analysis, although not applied on singular tasks. However, other researchers have applied the same or similar type of analysis on task level, for example Nyström (2008) and Bergqvist (2009), who both examine for what types of tasks students' performance can be explained to a relatively larger degree by students' reading ability.

Bergqvist (2009) and Roe and Taube (2006) have done their analyses on the same data, PISA 2003, opening up for the possibility to compare their results, which come from different methods. No in-depth comparisons are done here, but it can be noted that there are some similar results, for example that tasks with a higher demand of reading ability are to a higher extent tasks demanding constructed responses while tasks with a lower demand of reading ability to a higher extent are tasks of multiple-choice type (a result also replicated in the analysis of another data set by Nortvedt, 2009). This common result highlights the issue whether it really is a demand of *reading* ability that is measured with these methods, and not a demand of *writing* ability. However, when directly comparing issues of reading and writing, empirical results show that aspects of reading are more important (Österholm & Bergqvist, in press). There are also differences in the results from Bergqvist and Roe and Taube, for example regarding the demands of reading ability of tasks testing different processes or competencies (i.e. connections, reflections, and reproduction). Such differences highlight the issue of validity and reliability of the methods used.

Ansley and Forsyth (1990) do not use regression analysis but another method when examining the effect of the two abilities in relation to students' performance on mathematics tasks. They classify each student according to his/her level of reading ability (very low, low, high, very high) and according to his/her level of computation ability (very low, low, high, very high). Based on this classification, the students are divided into 16 groups, when put together in a 4 x 4 matrix. A separate matrix is created for each mathematics test item and in each cell in the matrix, the p-value (i.e. the proportion of students who answered the item correctly) for that particular group of students is calculated, see table 1. The p-values for the four extreme groups (very high or very low in both abilities) are denoted p_{HH}, p_{HL}, p_{LH}, and p_{LL} where the first index denotes level of reading ability and the second index denotes level of mathematical ability, and focus is on two differences; $p_{HL} - p_{LL}$ and $p_{LH} - p_{LL}$. Based on these differences, a task is characterized as (a) unidimensional if only one of these differences is large, (b) compensatory if both are large, and (c) noncompensatory if neither is large. Unidimensional tasks are seen as measuring primarily one ability, while for compensatory tasks the student can utilize either ability in order to solve the task (i.e. the student can compensate a deficiency in one ability with proficiency in another ability). Neither of the two abilities seems useful for solving noncompensatory tasks. Ansley and Forsyth's conclusions do not focus on making further characterizations of these three different types of tasks, but their focus is on a test as a whole for which they conclude that all these different types of tasks exist in all tests examined. They also discuss how it is possible to take these differences into consideration when constructing tests, that is, aspects of validity of written tests are discussed.

Table 1. *Basis for the analysis in the matrix method, using a 4 x 4 matrix.*

Reading ability	**Mathematical ability**			
	Very low	Low	High	Very high
Very low	p_{LL}			p_{LH}
Low				
High				
Very high	p_{HL}			

Walker et al. (2008) also discuss aspects of test validity but use another method when examining the effect of and relationships between different abilities. They use a multidimensional IRT analysis (item response theory), with mathematics tasks assumed to measure a pure mathematical ability (what they call ”naked” calculation tasks) as one dimension and specific tasks for testing reading ability as another dimension. Other mathematics tasks are then assumed to measure both these dimensions (abilities) and through the analysis a measure is created describing how close a task is to the second dimension (reading ability), giving a measure of a task's demand of reading ability. A DIF analysis (differential item functioning) then showed that for tasks with a high demand of reading ability, students with lower reading ability did indeed perform worse compared to other students, despite similar mathematical ability. The result from the DIF analysis can be interpreted as a validation of the IRT method in creating a measure of a task's demand of reading ability.

Since we do not use the same type of data as Walker et al. (2008) it is impossible for us to replicate the same type of analysis. Instead, we introduce another type of method for characterizing mathematics test items regarding their demands of reading ability; a principal component analysis. We have not found this method being used in other studies focusing on the relationship between reading and solving mathematical tasks.

We use four main methods of analysis from this area of research in our analysis of reliability and validity; (a) using correlations to reading ability, (b) using a regression analysis including measures of both reading ability and mathematical ability, (c) comparing p-values in the matrix method, and (d) using a principal component analysis. All these methods are described in more detail in the next section.

Method

Our main focus in this paper is to analyze aspects of validity and reliability of methods used to characterize mathematical test items regarding their demand of reading ability. We analyze aspects of validity based on the model of two types of abilities described earlier, regarding how well a method uses measures of these abilities in relation to the essential properties of homogeneity, separation, and overlap. Thus, in our discussions, we focus on certain aspects of construct validity. We see these aspects of validity as essential since the focus is on methods intended to examine the genuine effect of reading ability, which puts relationships between the two abilities at the core of our investigations. We analyze aspects of reliability by using data from PISA where we have access to results from test situations from different years where partly the same test items have been used. The test of reliability thus focuses on whether a method characterizes the mathematics test items in the same manner at two different test occasions, in 2003 and 2006.

Note that the analysis of reliability in this paper presumes that the demand of reading ability for a task can be described as a general property of the task. However, it could be that the demand of reading ability is dependent on mathematical ability, that there is a relevant type of interaction between these two variables in relation to their effects on student performance on mathematics tasks. Some of the methods examined also have the potential to study such interactions, but in this paper we do not perform such types of analyses. Instead we focus on a more general characterization of a task with respect to its demand of reading ability.

The assumption that the demand of reading ability is a general property of a task can be interpreted as seeing this property as a latent variable for the task, in the same manner as an ability is seen as a latent variable for a student. Such a variable can then be measured (i.e. approximated) using different methods. The validity and reliability of such methods are the focus of analyses in the present paper. The choice of data can be crucial to these analyses, in particular regarding which types of test items are used. For example, perhaps different methods would be preferred for different types of mathematics test items. Therefore, it would be valuable to repeat the type of analyses described in the present paper using other types of data.

The choice to use data from PISA in the present paper is partly based on the potential to examine aspects of reliability, as already mentioned. Another reason is that the set of data is large, both regarding the number of students and also regarding the number of test items. In addition, the students in PISA have completed test items in both mathematics and reading, which is suitable for our purpose of studying the relationships between abilities in these areas.

From the PISA data we only use results from Swedish students since otherwise the translation of test items becomes a factor. There are a total of 4624 and 4443 Swedish students in PISA 2003 and 2006 respectively. However, all students have not worked on all test items. There are results from around 1440 students per mathematics task in 2003 and around 1370 students per mathematics task in 2006. In addition, there is a large group of students who have not completed any reading test items at all. In our analysis, these students do not get a measure of reading ability (see below for more about measures of abilities). Therefore, those analyses that rely on an explicit measure of reading ability (i.e. methods 1–3 as described below) are based on even fewer students; depending on which task is analyzed, 340–740 students in 2003 and 670–700 students in 2006.

In PISA 2003 there are a total of 84 mathematics tasks and 28 reading tasks while in PISA 2006 there are a total of 48 mathematics tasks and 28 reading tasks. All tasks that were used 2006 were also used 2003. Thus, there are 48 mathematics tasks used both years, and these tasks constitute the basis for our analysis of reliability.

As described in the previous section, four different statistical methods are analyzed. These methods are described more below; complementing what is described in the previous section. The statistical tools used are suitable for the ordinal type of data in the results on single test items (i.e. 0, 1, 2 etc. points). The specific tools for each method are described below. In all analyses of statistical significance we use the significance level 0.05, but we also note when an analysis is significant at level 0.01.

Since three of the four methods rely on explicit measures of abilities, we first describe the method of creating these measures and thereafter we describe each of the four methods.

Measures of abilities

In the PISA database there are measures of abilities, the plausible values. However, we choose not to use these for our analyses, since plausible values are created in order to have good methods to estimate population parameters (Wu, 2005) and primarily not for analysis of singular students or singular test items, which is our main interest. When creating the plausible values, for example for mathematics, these values are not only based on students' results on mathematics test items but other information collected are also used. In particular, even students who have not completed any mathematics tasks at all do get plausible values for mathematics.

In accordance with our model of the two abilities, regarding the property of separation, we want to have as "pure" measures as possible of the two abilities. Therefore, we create our own measures of the abilities, but use the same general methodology as PISA; an IRT analysis (item response theory, see Hambleton, Swaminathan & Rogers, 1991), which is suitable for the data in PISA where all students have not worked on all tasks. However, as a contrast to PISA's plausible values, we only use mathematics tasks in the IRT analysis when creating measures of students' mathematical abilities and only use reading tasks in the IRT analysis when creating measures of students' reading abilities. Through the software Parscale we use a combination of 2- and 3-parametric partial credit models in our IRT analyses (three parameters are used for multiple-choice tasks), together with a Warm likelihood estimation (WLE) for abilities.

In order to test the validity of the ability measures we create, the correlations to plausible values are examined; taking the average of all 5 plausible values given for each ability in the PISA database. Even if we, as previously argued, see plausible values as less suitable for our specific purposes, there should still be a significant correlation between our measures and the plausible values. For the data in PISA 2003, the Pearson's correlation coefficient is very high between the different measures of mathematical ability ($r = 0.92$, $N = 4623$, $p < 0.001$) as it is also between the different measures of reading ability ($r = 0.91$, $N = 2458$, $p < 0.001$).

Method 1: Correlation with reading ability

The value of the correlation coefficient for the correlation between results on a single mathematics task and the measure of reading ability is here taken as a measure of the demand of reading ability for this task. Based on this measure of demand of reading ability, the tasks are ranked from having most (ranking 1) to least (ranking 48) demand of reading ability. A high correlation between the rankings in 2003 and 2006 is taken as a sign of good reliability. When Roe and Taube (2006) used this method they characterized a task as having high demand of reading ability if its correlation coefficient was higher than 0.4. As the authors admit, this limit is arbitrarily chosen, and in this paper we also test the reliability of this classification of tasks; whether tasks are classified as having high demand of reading ability both 2003 and 2006.

In this method we use a non-parametric tool for calculating correlation coefficients; Spearman's rank correlation coefficient.

Method 2: Regression

The value of the regression coefficient for the reading ability variable can be used in the same manner as the correlation coefficient in method 1 in order to rank tasks regarding their demand of reading ability and test the reliability of this ranking. In addition, the statistical significances of regression coefficients are also given when doing a regression analysis. Statistical significances create a classification of tasks; whether a task has reading ability and/or mathematical ability as significant in the regression model or not. Using notions from Ansley and Forsyth (1990), tasks are then classified as (a) unidimensional if only one of the coefficients is significant, (b) compensatory if both are significant, and (c) noncompensatory if neither is significant. Note that the category unidimensional actually consists of two subcategories; unidimensional with respect to mathematics and unidimensional with respect to reading. Further subcategories are also created if there are tasks with positive and negative regression coefficients. The reliability of this classification is tested in our analysis.

In this method we use ordinal logistic regression, suitable for the ordinal type of data we have for the dependent variables. In addition, the type of regression is chosen as *direct* logistic regression, since this method ”allows evaluation of the contribution made by each predictor over and above that of the other predictors" (Tabachnick & Fidell, 2006, p. 454), which is interpreted as the genuine effect of each of the predictors (i.e. mathematics and reading ability).

Method 3: The matrix method

The original matrix method, as described by Ansley and Forsyth (1990), relies on differences of p-values within the groups of students with lowest results of reading ability or mathematical ability. As the authors mention, the limit for when the differences are seen as large, is arbitrarily chosen. Instead of relying on this arbitrary limit, it is possible to use a test of statistical significance of the differences as a method for deciding if a difference is large or not. Since such a test also takes into consideration the number of students in the groups that are compared, we see this method as more suitable. However, in the PISA data there are relatively few students in some of the groups that are compared; the highest-lowest and lowest-highest groups, where for some tasks there is only one student in such a group. Instead of comparing these groups, an option is to study correlations within the groups with lowest abilities, where significant correlations are taken as the basis for characterizing tasks, instead of differences between p-values. That is, for the group with lowest mathematical ability, the correlation between reading ability and results on a specific task is calculated, and for the group with lowest reading ability, the correlation between mathematical ability and results on the task is calculated. Based on these correlations, tasks are classified similarly as in method 2; as (a) unidimensional if only one of the correlations is significant, (b) compensatory if both are significant, and (c) noncompensatory if neither is significant. As before, the category unidimensional consists of two subcategories; unidimensional with respect to mathematics and unidimensional with respect to reading, and further subcategories are also created if there are tasks with positive and negative correlation coefficients. The reliability of this classification is tested by examining if tasks are characterized in the same manner 2003 and 2006.

When using correlations in this method it is possible to create a ranking of mathematics tasks, in the same way as done in method 1, based on the correlation between reading ability and performance on a task, although now calculated only in a subgroup of students, where the variance of mathematical ability has been reduced. The reliability of this ranking is tested in our analysis.

Originally, Ansley and Forsyth (1990) used a 4x4 matrix for this method, but since this choice is also arbitrary, the method is here also tested using different sizes of the matrix, from 3x3 to 6x6, in order to examine the effects this might have on the reliability. In this method we use a non-parametric tool for calculating correlation coefficients; Spearman's rank correlation coefficient.

Method 4: Principal component analysis

In this method, no explicit measures of abilities are used, but the results on all test items, both mathematics and reading, are put into a principal component analysis. From this analysis it is possible to extract different components (dimensions) among the tasks, that is, those tasks that seem to measure "the same thing" are gathered in one component. Since we rely on the model of two types of abilities, we decide a priori to extract the first two components from this analysis. We expect the reading tasks to be placed in one component and at least most of the mathematics tasks in the other component. Some mathematics tasks could be placed in the reading component, which is then interpreted as showing a high demand of reading ability for these tasks.

In this method a principal component analysis with oblique rotation (since the components are expected to correlate) is performed. From the analysis, each mathematics task receives a loading value for each of the two components. The loading value on the reading component is taken as a measure of the demand of reading ability, which creates a ranking of tasks used to test the reliability. The limit of 0.32 for the absolute value of loading values is sometimes recommended to use for deciding if a task belongs to a component, although the limit is "a matter of researcher preference" (Tabachnick & Fidell, 2006, p. 649). Using this limit creates a classification of tasks similar to the classification in methods 2 and 3; that a task is (a) unidimensional if only one of the loading values is above the limit, (b) compensatory if both are above the limit, and (c) noncompensatory if neither is above the limit. As before, the category unidimensional includes two subcategories; whether the unidimensionality is with respect to mathematics or reading, and further subcategories are also created if there are tasks with positive and negative loadings. The reliability of this classification is tested by examining if tasks are characterized in the same manner 2003 and 2006.

Results and analysis

Aspects of validity

We have already mentioned one aspect of lack of validity for method 1 (correlation); that this method only takes into consideration reading ability and not mathematical ability. Since there is always a relatively strong correlation between these abilities[1], which is the basis for the overlap property, it is uncertain whether this method actually characterizes tasks specifically regarding their demands of reading ability.

Method 2 (regression) takes into account both reading and mathematical ability. An output from the analysis is how much of the variation of student performance on a specific task that can be explained by variation of reading ability, when also accounting for mathematical ability. Thus, there is good validity in this method regarding its ability to examine the genuine effect of reading ability on performance on a mathematics task.

Similarly as a regression analysis, the idea of method 3 (matrix) is to keep one variable (relatively) constant and then study the effect of the other variable. Making each student group smaller then keeps one variable relatively more constant (i.e. reducing the variation of this variable in the group), which could be seen as strengthening the validity of this method. The problem is the relatively strong correlation that exists between the two variables; reading and mathematical ability. Due to this correlation the variation of both variables are reduced when creating smaller student groups. As described in the method section, this problem is evident already when using 4 groups since for some tasks there is only one student having lowest ability regarding one variable and highest ability regarding the other variable.

There is also another problem with method 3. Despite reducing the variation in mathematical ability within a group of students, for many tasks there is still a significant correlation between mathematical ability and student performance on the task. This is true for almost all tasks when using the 3 x 3 matrix and about half of the tasks when using the 6 x 6 matrix. Thus, within the group with lowest mathematical ability the effect of mathematical ability on performance was not excluded, and an observed effect of reading ability in this group can therefore come from the overlap between the two abilities.

Methods 2 and 3 share a common problem, caused by their reliance on the measure of mathematical ability. There is a kind of dilemma since on the one hand the measure is assumed to be purely about mathematics, and not reading, and on the other hand it is acknowledged that some mathematics tasks used when creating the measure put (perhaps high) demands on reading ability. Thus, for these methods, the separation between the two abilities is not ideal.

Since no measure of ability is used in method 4 (principal component analysis), the same problem as noted for methods 2 and 3 about the separation of the different abilities does not exist in method 4. Instead, the principal component analysis is a more bottom-up type of analysis through which we analyze the results on all test items in order to find out which items have most in common, which is assumed to reflect that these items measure something similar. For the PISA data, both 2003 and 2006, the principal component analysis yielded an anticipated structure with reading tasks clearly gathering on one component while mathematics tasks tend to gather on another component (see table 2). This result is seen as a sign of good validity for this method in relation to the model with two abilities, in particular regarding separation and also, but to lesser degree, regarding homogeneity. Due to the relatively high proportion of mathematics tasks with no high loading on either of the first two components (see table 2), for future studies and methodological developments, it could be of interest to examine a more homogenous set of mathematics tasks or to examine the use of more than two components in this method.

Table 2. *Proportion of tasks with a loading value larger than 0.32 on extracted components in method 4 (no task has high loadings on both components).*

Test items	**Component**		
	Math	Reading	None
Math 2003 ($n=84$)	49 %	24 %	27 %
Reading 2003 ($n=28$)	0 %	93 %	7 %
Math 2006 ($n=48$)	56 %	4 %	40 %
Reading 2006 ($n=28$)	0 %	89 %	11 %

Methods 2, 3 and 4 are similar regarding their potential to give more in-depth information in the classification about the relationships between reading and mathematical abilities when solving mathematics tasks. Tasks are not only characterized as having or not having high demands of reading ability, but the categories unidimensional (including two subcategories), compensatory, and noncompensatory are used. This property of the methods can be seen as a strength regarding validity since the results from analyses then can be seen as more fully describing the relationships between the two abilities. However, for the regression analysis (method 2) this property of the method became only a *theoretical* potential, since all tasks in our data had a significant genuine demand of mathematical ability, in practice reducing the number of categories to two; unidimensional regarding mathematics and compensatory. It is uncertain if this limitation is a more general property of this method when trying to characterize the demands of reading ability of mathematics tasks or if this limitation is more specific to the type of data used in this study. Unlike the regression analysis, the uses of the matrix method and the principal component analysis (methods 3 and 4) resulted in tasks classified as unidimensional regarding reading. However, in the matrix method these tasks are rather few and are only present for matrix sizes 4 x 4 and 5 x 5, when less than three tasks have this property. In the principal component analysis there are more tasks classified as unidimensional regarding reading in 2003 (8 tasks) but fewer in 2006 (2 tasks).

It should also be noted that this more precise characterization of the demands of reading ability cannot directly be interpreted in the same way in all three methods. In the regression analysis, only genuine effects of abilities are examined, where effects of the overlap between the two abilities are excluded. The property of the principal component analysis is similar, when the components are seen as representing the abilities, since the loading values used (from the pattern matrix) ”represent the unique contribution of each factor [component] to the variance of each variable but do not include segments of variance that come from overlap between correlated factors [components]” (Tabachnick & Fidell, 2006, p. 627). While the intention is the same for the matrix method, that is, to study the genuine effect of reading ability, the problem of an existing correlation between mathematical ability and performance also in the group with lowest mathematical ability causes the two abilities to be mixed in an unwanted manner.

In conclusion, the largest lacks of validity are noted for methods 1 and 3, when using correlations and the matrix method. No fundamental problems are noted for methods 2 and 4, when using regression analysis and principal component analysis, but there is some uncertainty in method 2 regarding the separation between the abilities, caused by the creation of explicit measures of the abilities.

Aspects of reliability

Within the four methods there are two different ways to characterize mathematics tasks regarding their demands of reading ability; one measure that creates a ranking of the tasks and one classification of tasks into different categories. Table 3 summarizes how these two ways of characterizing tasks are used in the four methods together with results from the analysis of reliability. The theoretical measurement for classification in methods 2, 3 and 4 is identical; using the categories unidimensional, compensatory, and noncompensatory, together with subcategories caused by positive and negative coefficients or loadings. However, as also mentioned in the discussion of validity, for the data analyzed in this study, some categories do not contain any tasks. A priori, we did not expect any negative coefficients or loadings to appear, but in the regression analysis (method 2) some tasks (3 in 2003 and 3 in 2006) have a significant negative coefficient for reading ability. However, in methods 3 and 4 (matrix and principal component analysis) no task has significant negative coefficient or loading. In table 3, regarding measurement for classification, only the categories that are non-empty for each method are listed.

Table 3. *Summary of analysis of reliability for the four methods based on a total of 48 mathematics tasks (except for method 3 where the analysis is based on fewer tasks, as described below).*

Method	**Measurement for:**		**Reliability of:**	
	Ranking	Classification	Ranking[1]	Classification[2]
1	Correlation coefficient	a) Coefficient > 0.4 b) Coefficient < 0.4	0.79^{**}	87.5 %
2	Regression coefficient	a) Signif. positive read coeff. b) Signif. negative read coeff. c) Non-signif. read coeff.	0.18	62.5 %
3 (3x3)[3]	Correlation coefficient	a) Only math coeff. significant	0.41^{**}	60.0 %
3 (4x4)[4]		b) Only read coeff. significant	0.26	46.5 %
3 (5x5)[4]		c) Both coefficients significant	0.05	46.5 %
3 (6x6)[5]		d) No coefficient significant	-0.11	37.5 %
4	Loading value	a) Only math loading > 0.32 b) Only read loading > 0.32 c) Both loadings > 0.32 d) No loading > 0.32	0.75^{**}	70.8 %
		a) Read loading > math loading b) Read loading < math loading		87.5 %

[1] The value gives the correlation between ranking 2003 and 2006, with statistical significance marked; $^{*}p < 0.05$ and $^{**}p < 0.01$.

[2] The value gives the proportion of tasks classified the same way in 2003 and 2006.

[3] Analysis based on 45 tasks.

[4] Analysis based on 43 tasks.

[5] Analysis based on 40 tasks.

Regarding the ranking of tasks described in table 3, the reliability is analyzed by examining the correlation between the measures from

2003 and 2006, in order to see to what degree a similar kind of ranking is created both years. Regarding the classification of tasks, the reliability is analyzed by calculating the percentage of tasks that have been categorized in the same way both 2003 and 2006.

When using the matrix method there are some situations where no correlation coefficient can be computed due to lacking variability in the data for a group of students for a specific task, in particular that for some tasks all students have received the same score. Therefore, this method cannot be applied on all test items. This type of problem can in itself be seen as a lack of reliability for this method, and as a consequence the analysis of reliability is not always based on 48 tasks for this method, as described in table 3.

There is a strong reliability of the ranking, also of similar magnitude, for methods 1 and 4, while the reliability of the classification is not as high for method 4 as for method 1. However, the classification used in method 4 is more complex than the one used in method 1. Therefore, in order to make the classifications used in these methods more directly comparable, a more simple type of classification is also tested for method 4, as included in table 3. This type of classification in method 4 yields similar magnitude of reliability as in method 1.

A clear lack of reliability is noted for methods 2 and 3. For the matrix method the reliability is lower when making the student groups smaller (i.e. when using larger matrices). Thus, methods 1 and 4 show the best reliability, with similarly high values regarding both the ranking and the classification, at least when using comparable (and simpler) types of classifications.

Conclusions and discussion

Among the four methods examined, our conclusion is that the principal component analysis (method 4) has the best properties when taking into account aspects of both validity and reliability. Thus, if the purpose is to quantitatively characterize mathematics tasks regarding their demands of reading ability and the data includes students' results on tasks for both mathematics and reading, a principal component analysis is preferred.

There are many ways to refine and develop the types of analysis focused on in this paper, regarding both empirical methods and also connections to theories and models about effects of, and relations between, different abilities on the performance on mathematics tasks.

Regarding the overlap between the abilities in mathematics and reading, we have argued for disregarding the effect of this overlap and focused on the genuine effect of reading ability. Even if it is statistically quite easy to distinguish between the common effect of the two abilities and a genuine effect of one ability, there is a need to connect this type of analysis to theories or models regarding the relationship between different abilities when solving mathematical tasks. Furthermore, the overlap between abilities can have different causes and to examine properties of this overlap could add new knowledge about the relationships between the abilities.

Except the common effect of two abilities described by the overlap between them, there is also the potential interaction between variables to take into account. In particular, an interaction exists when the demand of reading ability for a task is not independent of mathematical ability. Such interactions can be studied directly in some of the methods examined in this paper, in particular when using a regression analysis or the matrix method. However, since our analysis has shown some deficiencies in these methods there is a need for methodological development regarding the study of the interaction between reading ability and mathematical ability in relation to the solving of mathematics tasks. Furthermore, there are empirical results about the reading of longer texts showing an interaction between text complexity and student ability in relation to reading comprehension (McNamara et al., 1996). This type of interaction could therefore be of interest to examine also for the reading of test items.

Some of the deficiencies regarding reliability and validity might not have any large practical effect if the focus is on the characterization of the group of tasks with highest demand of reading ability, rather than on single tasks. This potential robustness of methods is perhaps observed when some similar results are noted by both Bergqvist (2009) and Roe and Taube (2006) despite noted lack of reliability in the method used by Bergqvist and lack of validity in the method used by Roe and Taube. However, it seems more suitable to use the method with better reliability and validity, the principal component analysis, also since this method is somewhat simpler than the other methods because it does not rely on any explicit measures of abilities, which can be complex to create.

Focus in this paper is on studies in research area 3 (as labeled in the background), regarding statistical analyses of effects of different abilities on students' performance when solving mathematical tasks. In the background we also criticize several studies in research areas 1 and 2, regarding studies about effects of linguistic properties of mathematical tasks and about the process of reading and solving mathematical tasks. However, we see it as essential to combine foci from different research areas in order to better understand the connections between reading and solving mathematical tasks.

In future quantitative studies we will use the results from a characterization of mathematics tasks regarding their demands of reading ability together with other types of characterizations of the same tasks, in order to combine the foci in research areas 1 and 3. For example, we have started to examine if there are any special linguistic properties of tasks having a high demand of reading ability (Bergqvist et al., in press; Österholm & Bergqvist, in press).

We will also include more qualitative types of studies in our future attempts to deepen our understanding of the relationships between reading and solving mathematical tasks. For example, we intend to include both a student perspective, regarding their view of tasks with different demands of reading ability and what strategies are used when solving these tasks, and also a teacher perspective, regarding their view of, and in teaching use of, tasks with different demands of reading ability. These types of studies are meant to combine the foci in research areas 2 and 3.

Finally, our ambitions are also to continuously relate results from empirical studies to the development of theories and models about the relationships between reading and solving mathematical tasks.

References

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

Aiken, L. R. (1972). Language factors in learning mathematics. *Review of Educational Research*, 42, 359–385.

Ansley, T. N. & Forsyth, R. A. (1990). An investigation of the nature of the interaction of reading and computational abilities in solving mathematics word problems. *Applied Measurement in Education*, 3 (4), 319–329.

Bergqvist, E. (2009). A verbal factor in the PISA 2003 mathematics items: tentative analyses. In M. Tzekaki, M. Kaldrimidou & C. Sakonidis (Eds.), *Proceedings of the 33rd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 145–152). Thessaloniki: PME.

Bergqvist, E., Dyrvold, A. & Österholm, M. (in press). Relating vocabulary in mathematical tasks to aspects of reading and solving. In *Proceedings of the Eighth Swedish Mathematics Education Research Seminar, Madif 8.*

Bergqvist, E. & Österholm, M. (2010). A theoretical model of the connection between the process of reading and the process of solving mathematical tasks. In C. Bergsten, E. Jablonka & T. Wedege (Eds.), *Mathematics and mathematics education: Cultural and social dimensions* (Proceedings of MADIF 7) (pp. 47–57). Linköping: SMDF. Retrieved December 14, 2010, from http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-31890

Bernardo, A. B. I. (1999). Overcoming obstacles to understanding and solving word problems in mathematics. *Educational Psychology*, 19 (2), 149–163.

Burton, L. & Morgan, C. (2000). Mathematicians writing. *Journal for Research in Mathematics Education*, 31, 429–453.

Cowen, C. C. (1991). Teaching and testing mathematics reading. *American Mathematical Monthly*, 98 (1), 50–53.

Fuentes, P. (1998). Reading comprehension in mathematics. *Clearing House*, 72 (2), 81–88.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage Publications.

Homan, S., Hewitt, M. & Linder, J. (1994). The development and validation of a formula for measuring single-sentence test item readability. *Journal of Educational Measurement*, 31 (4), 349–358.

Knifong, J. D. & Holtan, B. D. (1977). A search for reading difficulties among erred word problems. *Journal for Research in Mathematics Education*, 8 (3), 227–230.

Konior, J. (1993). Research into the construction of mathematical texts. *Educational Studies in Mathematics*, 24, 251–256.

Krygowska, Z. (1969). Le texte mathématique dans l'enseignement. *Educational Studies in Mathematics*, 2, 360–370.

Lager, C. A. (2006). Types of mathematics-language reading interactions that unnecessarily hinder algebra learning and assessment. *Reading Psychology*, 27, 165–204.

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

McNamara, D. S., Kintsch, E., Songer, N. B. & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, 14, 1–43.

Muth, K. D. (1984). Solving arithmetic word problems: Role of reading and computational skills. *Journal of Educational Psychology*, 76 (2), 205–210.

Möllehed, E. (2001). *Problemlösning i matematik: en studie av påverkansfaktorer i årskurserna 4–9*. Malmö: Institutionen för pedagogik, Lärarhögskolan.

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

Niss, M. & Jensen, T. H. (Eds.). (2002). *Kompetencer og matematiklæring – idéer og inspiration til udvikling af matematikundervisning i Danmark*. København: Undervisningsministeriets forlag. Retrieved August 6, 2008, from http://pub.uvm.dk/2002/kom/hel.pdf

Nortvedt, G. A. (2009). The relationship between reading comprehension and numeracy among Norwegian grade 8 students. In M. Tzekaki, M. Kaldrimidou & H. Sakonidis (Eds.), *Proceedings of the 33rd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 233–240). Thessaloniki, Greece: PME.

Nyström, P. (2008). *Identification and analysis of text-structure and wording in TIMSS-items*. Paper presented at the 3rd IEA International Research Conference. Retrieved August 23, 2009, from http://www.iea.nl/fileadmin/user_upload/IRC2008/Papers/TIMSS_Mathematics/Nystrom.pdf

Oakland, T. & Lane, H. B. (2004). Language, reading, and readability formulas: Implications for developing and adapting tests. *International Journal of Testing*, 4 (3), 239–252.

OECD. (2003). *The PISA 2003 assessment framework – mathematics, reading, science and problem solving knowlegde and skills*. Paris: Author.

OECD. (2006). *Assessing scientific, reading and mathematical literacy: a framework for PISA 2006*. Paris: Author.

Paul, D. J., Nibbelink, W. H. & Hoover, H. D. (1986). The effects of adjusting readability on the difficulty of mathematics story problems. *Journal for Research in Mathematics Education*, 17, 163–171.

Roe, A. & Taube, K. (2006). How can reading abilities explain differences in maths performance? In J. Mejding & A. Roe (Eds.), *Northern lights on PISA 2003 – a reflection from the Nordic countries* (pp. 129–141). Copenhagen: Nordic Council of Ministers. Retrieved August 6, 2008, from http://www.norden.org/pub/uddannelse/uddannelse/sk/TN2006523.pdf

Sato, E., Rabinowitz, S., Gallagher, C. & Huang, C.-W. (2010). *Accommodations for English language learner students: the effect of linguistic modification of math test item sets*. Washington: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved October 13, 2011, from http://ies.ed.gov/ncee/edlabs/regions/west/pdf/REL_20094079.pdf

Shanahan, T. & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: rethinking content–area literacy. *Harvard Educational Review*, 78 (1), 40–59.

Søvik, N., Frostad, P. & Heggberget, M. (1999). The relation between reading comprehension and task-specific strategies used in arithmetical word problems. *Scandinavian Journal of Educational Research*, 43 (4), 371–398.

Tabachnick, B. G. & Fidell, L. S. (2006). *Using multivariate statistics* (5th ed.). Boston: Allyn and Bacon.

Walker, C. M., Zhang, B. & Surber, J. (2008). Using a multidimensional differential item functioning framework to determine if reading ability affects student performance in mathematics. *Applied Measurement in Education*, 21, 162–181.

Vinner, S. (1997). The pseudo-conceptual and the pseudo-analytical thought processes in mathematics learning. *Educational Studies in Mathematics*, 34, 97–129.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, 31, 114–128.

Österholm, M. (2007). A reading comprehension perspective on problem solving. In C. Bergsten & B. Grevholm (Eds.), *Developing and researching quality in mathematics teaching and learning* (Proceedings of MADIF 5) (pp. 136–145). Linköping: SMDF. Retrieved June 7, 2012, from http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-14116

Österholm, M. & Bergqvist, E. (in press). What mathematical task properties can cause an unnecessary demand of reading ability? In G. H. Gunnarsdóttir, F. Hreinsdóttir, G. Pálsdóttir, M. Hannula, M. Hannula-Sormunen, et al. (Eds.), *Proceedings of Norma 11, The sixth Nordic conference on mathematics education in Reykjavík, May 11–14, 2011* (pp. 661-670). Reykjavík: University of Iceland Press.

Notes

[1] In the background we describe studies showing significant correlations between 0.40 and 0.86. From the data used in the present study, there are significant correlations of 0.57 and 0.58 from the year 2003 and 2006 respectively.

Magnus Österholm

Magnus Österholm has a PhD in mathematics education from Linköping University and now works as a research fellow at the Department of Science and Mathematics Education at Umeå University. He is also a member of Umeå Mathematics Education Research Centre (UMERC). During 2011 and 2012 he is a visiting scholar at Monash University in Melbourne, Australia. His research interests deal primarily with mathematics education at the upper secondary and university levels, where cognitive and metacognitive perspectives are of special interest, together with studying language and communication in the learning and teaching of mathematics.

magnus.osterholm@matnv.umu.se

Ewa Bergqvist

Ewa Bergqvist has a PhD in mathematics education from Umeå University and is an assistant professor at the Department of Science and Mathematics Education at Umeå University. She is a member of Umeå Mathematics Education Research Centre (UMERC) and a teacher in mathematics education for pre-service mathematics teachers. Her research focuses mainly on language, competencies, and reasoning in upper secondary and university level mathematics.

ewa.bergqvist@matnv.umu.se