WHAT MATHEMATICAL TASK PROPERTIES CAN CAUSE AN UNNECESSARY DEMAND OF READING ABILITY?

Magnus Österholm and Ewa Bergqvist
Umeå University, Sweden

In this study we utilize test results from Swedish students in PISA 2003 and 2006 to examine what types of task properties predict the demand of reading ability of a mathematical task. In particular, readability properties (sentence length, word length, common words, and information density) and task type properties (content, competence, and format) are examined. The results show that it is primarily readability properties of a task that predict the task's demand of reading ability, in particular word length and to some extent information density (measured through the noun-verb quotient).

Keywords: language, readability, PISA, assessment, validity

INTRODUCTION

When discussing validity of mathematics test items, the complexity of the text is sometimes highlighted. For example, in the PISA studies it is stated that the mathematics test should use as simple language as possible (OECD, 2006), in order to avoid that mathematics test items measure reading ability instead of mathematical ability. However, it is not trivial to decide exactly what can be regarded as a purely linguistic simplification of a test item and what linguistic properties of a test item are also parts of the mathematical content of the task. For example, in one study, item texts were altered in some word problems so that the relationship between known and unknown quantities became more explicit, which the author describes as a linguistic change (Bernardo, 1999). However, this type of alteration could perhaps also be seen as affecting the validity of the task regarding its potential in testing a student's mathematical ability, e.g. regarding a competence of mathematical modelling.

A linguistic complexity of a task text is not necessarily a sign of low validity regarding the mathematical content and competence being tested or practiced with this task. In particular, a communication competence is often included in frameworks of mathematical knowledge (e.g. NCTM, 2000), where aspects of reading and writing can also be included. Therefore it can be difficult to locate and describe a border between types of communication that are seen as part of mathematical knowledge and types of communication that are not, which if included in a mathematical task can be seen as having a negative effect on the validity of the task.

Thus, there is a need for research that focus on relationships between reading and solving mathematical tasks, for example in order to decide what text properties of mathematical tasks can be altered in order to improve the validity of the tasks. In this paper we contribute to this endeavour by using a new type of method for the analyses of different task properties regarding if these properties can be said to cause an unnecessary demand of reading ability for students trying to solve the task.

BACKGROUND

There are plenty of studies investigating aspects of reading in relation to solving mathematical tasks, using many different methods. In our previous discussion about these methods (see Österholm & Bergqvist, in press), we notice that many of these methods do not in a satisfactory manner locate aspects of language or communication that primarily are not part of the mathematics (i.e. mathematical content or competence). For example, there are studies that in a direct way examine connections between different (linguistic) properties of tasks and students' performance on these tasks. From such studies it is difficult to decide if the examined properties are relevant or not in relation to the mathematics of the task.

One way to examine if some task properties are not mainly an aspect of the mathematics is to examine if these properties primarily create a higher demand of students' reading ability and not mathematical ability. Several studies take this approach by utilizing situations when students take both a test of reading ability and of mathematical ability, where the effect of these abilities on the solving of mathematical tasks can be examined. Different statistical methods are used in such studies, primarily utilizing correlations and regressions in different ways. We have previously examined these types of methods regarding their validity and reliability (Österholm & Bergqvist, in press), and the results show that methods we found in other studies all have problems with aspects of validity or reliability. The problem regarding validity consists of a failure to locate a genuine effect of reading ability, separated from mathematical ability, on students' solving of mathematical tasks. The problem regarding reliability consists of a failure to replicate a characterization of a mathematical task regarding its demand of reading ability when the same task is used with different groups of students. More detailed explanations and arguments concerning the validity and reliability of different methods can be found in Österholm and Bergqvist (in press). In our previous study we also introduced a type of statistical method for characterizing mathematical tasks regarding their demand of reading ability that we have not seen used in this way before. This method utilizes a form of factor analysis, and proved to have good properties regarding both validity and reliability (this method and its properties are described more in the method section). Therefore, we use this method in the present paper in order to characterize mathematical tasks regarding their demand of reading ability and then examine what types of properties of tasks can predict this demand of reading ability.

Although many methods have shown deficiencies regarding validity and reliability, as discussed above, studies using these methods show some similar results. In parti-

cular, it has been observed in several studies, which have used different methods and different data, that when students for a specific task need to produce a written answer and not only choose one of several given possible answers, this task is characterized as having a higher demand of reading ability (Bergqvist, 2009; Nortvedt, 2009; Roe & Taube, 2006). This common result questions if it is primarily the aspect of reading that is measured since the result highlights the aspect of producing a written answer. Perhaps it is a more general type of communicative ability that is measured, and where the writing ability appears as most relevant for mathematical tasks. As a part of the present study we examine the relationship between aspects of reading and writing.

Theoretical considerations

Aspects of reading ability and mathematical ability, and their relations to the solving of mathematical tasks are central as a basis for the methods of analysis utilized in the present paper. The basis for our analyses is a model where we include the two abilities (i.e. latent variables) seen as (the most) relevant for the outcome of solving mathematical tasks. Included in the model is an assumption that these two abilities are separated and homogenous enough in order to study them as two different dimensions of human cognition/abilities. The main interest in the present study is the genuine effect of reading ability on the solving of mathematical tasks. The magnitude of this genuine effect of reading ability on the outcome of a specific mathematical task is labelled as the task's demand of reading ability. More detailed discussions concerning relationships between aspects of theory and methods of analysis can be found in Österholm and Bergqvist (in press).

PURPOSE

The primary research question in this study is:

(Q1) What types of task properties can predict the demand of reading ability of a mathematical task?

In particular, our aim with this paper is to examine different *types* of task properties. On the one hand we examine properties that are included in different measures or formulas regarding readability, which are seen as connected to aspects of reading. These types of properties are here labelled as the *readability properties*. On the other hand we also examine properties based on classifications of tasks regarding mathematical content and competence, and also regarding what type of answer is asked for. These types of properties are here labelled as the *task type properties*. Properties regarding mathematical content and competence are included here mainly to try to confirm that these properties are *not* related to demand of reading ability. However, we are also interested in examining whether the demand of reading ability is primarily connected to properties regarding aspects of reading or to other types of properties, in particular aspects of writing. This interest is caused by the common empirical result from other studies, showing a connection between the demand of reading

ability and the demand of producing a written response (Bergqvist, 2009; Nortvedt, 2009; Roe & Taube, 2006). Thus, we have a second research question:

(Q2) Is a mathematical task's demand of reading ability mainly connected to a demand of reading (through task properties about readability) or to a demand of writing (through what type of answer is asked for)?

Compared to previous research about relationships between reading and solving mathematical tasks regarding what task properties are connected to a task's demand of reading ability, the contribution of the present study is manifold. Firstly, we use a form of factor analysis as method for examining a task's demand of reading ability, a method we have not seen used before but which has shown to have better validity and reliability compared to other methods (Österholm & Bergqvist, in press). Secondly, in our analysis we connect aspects of readability to the demand of reading ability and not directly to student performance, which we see as more suitable since many factors influence student performance and our method of analysis pinpoint properties that seem to cause an unnecessary linguistic complexity. Finally, we examine if aspects of reading or writing have the strongest effect on the demand of reading ability, adding to previous results that primarily seem to examine either aspects of reading or aspects of writing.

METHOD

In order to measure a mathematics task's demand of reading ability we use a principal component analysis (PCA), which we shortly refer to as a form of factor analysis, although usually not defined specifically as a *factor* analysis. All Swedish students' results on all PISA mathematics test items and reading test items from 2003 and 2006 are entered into the factor analysis. In this analysis we use Promax as the method for rotation (i.e. oblique rotation, since we expect the factors to correlate) and extract only the first two factors, which are expected to correspond to the two abilities of mathematics and reading. From this analysis, each mathematics item receives a loading value for each of the two factors. The loading value on the reading factor is taken as a measure of the demand of reading ability. In our previous study, this method was deemed to have good validity and reliability (Österholm & Bergqvist, in press): Regarding validity, this use of factor analysis created an anticipated division into two factors where almost all reading items were placed in one factor and the other factor had only mathematics items with high loading values. In addition, the loading values created through the factor analysis correspond to "the unique contribution of each factor to the variance of each variable but do not include segments of variance that come from overlap between correlated factors" (Tabachnick & Fidell, 2006, p. 627). That is, the loading value can be interpreted as a measure of the genuine effect of reading ability when the effect of mathematical ability has been excluded. Regarding the reliability of this method, the characterization of test items

regarding their demand of reading ability was very consistent when applied on test items included both in PISA 2003 and in PISA 2006.

The loading value from the factor analysis can be positive or negative, for which it is a *qualitative* difference since reading ability then either has a positive or negative effect on the solving of an item. Items having positive or negative loading values must therefore be analyzed separately. Since the purpose of this paper is to analyze predictions of the demand of reading ability, we focus on items with a positive loading value since these items indeed have a varying *demand* of reading ability.

The data used in this study (PISA 2003 and 2006) are suitable since they contain results on tests of mathematics and reading from the same students, and the analysis can be based on results from many students. For our analyses, reading and mathematical ability are thus defined as reading and mathematical literacy according to the PISA framework, which includes many types of competences in reading and mathematics (OECD, 2006). In future studies it could be of interest to use other definitions of reading ability and mathematical ability in similar empirical analyses.

Since all the reading tasks and many of the mathematics tasks were the same in PISA 2003 and 2006, we combine the results from these years in our analyses. Thereby, we have data from a total of 9067 Swedish students, but since all tasks were not given to all students there are results from fewer students on each task; around 2700 students for tasks used both years, and around 1400 for tasks used only in 2003. There are 84 mathematics tasks used in PISA 2003 and 2006, of which 63 have a positive loading from the factor analysis. These 63 items are the basis for our analyses in this paper.

Readability properties

In our analysis we do not use readability formulas that have been created through linear combinations of different types of text properties, e.g. as is the case in the Homan-Hewitt formula (Homan, Hewitt, & Linder, 1994) and in the Swedish readability index LIX (Björnsson, 1968). Instead of relating to this type of abstract measure of readability, we want to relate directly to specific properties of the texts, in order to examine what properties are connected to the demand of reading ability. Therefore, we use measures that in a more direct manner measure certain properties of texts, but where parts of more complex readability formulas can be utilized. In this paper we focus on the following properties and measures:

- Sentence length: Measured through the average number of words per sentence, which is used as part of the LIX formula (Björnsson, 1968).
- Word length: Measured through the fraction of long words, which are words longer than six letters, which is used as part of both the LIX formula (Björnsson, 1968) and also the Homan-Hewitt formula (Homan et al., 1994).
- Common words: Here measured as the fraction of words included in the 1000 most common words, according to some corpus. For this we use two Swedish corpora; SUC, which is a balanced corpus with one million words

(Gustafson-Capková & Hartmann, 2006), and OrdiL, which consists of words from common Swedish lower secondary textbooks (Lindberg & Johansson Kokkinakis, 2007). This type of property is part of the Homan-Hewitt formula (Homan et al., 1994).

- Information density: A complex type of property, but here measured through the noun-verb quotient ('nominalkvot' in Swedish, see Einarsson, 1978), which is the number of nouns divided by the number of verbs.

All measurements are calculated through a computerized analysis of the PISA mathematics item texts (see Liberg & Forsbom, 2009), thus creating high reliability in the calculations. However, there can be some problems regarding validity in creating quantitative measures of properties of test items since these texts are often very short and can include several semiotic systems, e.g. also formulas and diagrams, that can be difficult to handle in the calculations. In particular, regarding the noun-verb quotient, we get some very high values for some tasks, up to 17, while common values for "ordinary" written texts are between 0.9 and 1.7 (Einarsson, 1978, p. 97). We do not interpret these extreme values as signs of extreme high information density, but as a sign that this calculation is not a suitable measure for all task texts. For example, we see lists of nouns used in tables or diagrams as a potential cause of this unsuitableness. In our analyses, we choose to exclude outliers (according to the procedure described below) in order to avoid this problem with extreme values. We could use other means to avoid this problem, in particular to exclude some parts of item texts when calculating noun-verb quotients. However, more in-depth analysis is needed to create criteria for this, which is a topic for future studies.

There can also be some problems with the other measures. For sentence length, each part of a table counts as one sentence. Such a situation can create many short sentences, which can distort the measure of sentence length as an aspect of readability. For measures focusing on properties of singular words, a problem can be what to count as a word, e.g. regarding formulas. Thus, similarly as for the verb-noun quotient, it could be reasonable to exclude some parts of the item texts for the other measures. However, in this study no text is excluded when measuring readability properties, instead we use the exclusion of outliers as a means to avoid some potential problems when quantifying properties of item texts. Absolute values of z-scores higher than 3.29 are treated as outliers (Tabachnick & Fidell, 2006). Since we treat outliers as faulty values, after outliers have been removed, new z-scores are calculated among remaining values, and any new outliers are removed. This procedure is repeated until no more outliers exist. Using this procedure for the mathematics PISA items results in 57 items for analyses that include the verb-noun quotient, 62 items for sentence length, and all 63 items for analyses of word length and common words.

Task type properties

Categories from PISA's framework (OECD, 2006) are used as task type properties, and the categorization of the PISA tasks done by OECD is also used in the analyses:

- Overarching ideas (mathematical content): Uncertainty, Change and relationships, Space and shape, or Quantity.
- Competency cluster: Reproduction, Connections, or Reflection.
- Format (type of answer): Multiple choice (MC), Complex multiple choice (CMC), Closed constructed response (CCR), or Open constructed response (OCR).
- Format-written: Whether a written answer is needed (CCR and OCR) or not (MC and CMC).

Since our main interest regarding format is whether a written answer is needed or not the dichotomous Format-written variable is created as described above.

Statistical analyses

The first step of the statistical analyses is to examine what properties have significant connections to demand of reading ability (research question Q1). For the quantitative readability properties, correlations are used, while for the nominal task type properties, ANOVAs are used. For the dichotomous Format-written variable, a t-test is used. In the next step of the statistical analyses, we use regression analyses in order to examine the relationships between those properties that have a significant connection to demand of reading ability (research question Q2). In all statistical analyses, we use a significance level of 0.05. Parametric methods are used and reported, but corresponding non-parametric methods are also performed in order to test the stability. We report on the congruence or mismatch between these types of methods.

RESULTS

Table 1. Analysis of which properties are statistically connected to demand of reading ability, where r-values refer to Pearson correlation coefficients, F-values are from ANOVAs, and the t-value is from a t-test

Property	Items analyzed	Statistics	Significance
Sentence length	62	r = -0.008	p = 0.952
Word length	63	r = 0.318	p = 0.011
Common words, SUC	63	r = -0.195	p = 0.126
Common words, OrdiL	63	r = -0.193	p = 0.130
Information density	57	r = 0.339	p = 0.010
Overarching ideas	63	F(3,59) = 0.685	p = 0.565
Competency cluster	63	F(2,60) = 3.690	p = 0.031
Format	63	F(3,59) = 1.769	p = 0.163
Format-written	63	t(61) = 2.295	p = 0.025

The results in table 1 show that four different properties have a significant statistical connection to the demand of reading ability (p-values below 0.05). When corresponding non-parametric methods are used, the same statistical significances occur, except for the property of Competency cluster. Because of these contradictory results, we test whether conditions for an ANOVA are fulfilled, by using Levene's homogeneity-of-variance test. This test shows that there is a difference between the variances when analyzing the effect of Competency cluster ($F(2,60)= 3.59$, $p=0.034$). Thus, the conditions for an ANOVA are violated, and therefore we rely on the non-parametric test; that there is no significant connection between Competency cluster and demand of reading ability. Therefore, we focus on three properties in the next step of analysis; Word length, Information density, and Format-written. Since the noun-verb quotient is included, the analyses are based on 57 items. The results from a regression analysis using all three variables show that the regression model explains a significant proportion of variance in demand of reading ability, with $R^2=0.203$ ($F(3,53)=4.492$, $p=0.007$).

Table 2. Significance of added R^2 (i.e. explained variance) when entering a second independent variable to a regression analysis with demand of reading ability as dependent variable

	Second variable		
First variable	*Word length*	*Info density*	*Format-written*
Word length		$F(1,54) = 3.409$ $p = 0.070$	$F(1,54) = 1.977$ $p = 0.165$
Info density	$F(1,54) = 4.656$ $p = 0.035$		$F(1,54) = 2.143$ $p = 0.149$
Format-written	$F(1,54) = 5.933$ $p = 0.018$	$F(1,54) = 4.833$ $p = 0.032$	

Table 2 summarizes results from different sequential regression analyses with only two of the variables at a time. These results give information about the relationships between the variables regarding their effect on demand of reading ability:

- When Format-written is first entered and thereafter one of the readability properties, there is a significant additional effect of each readability property.
- When one of the readability properties is first entered and thereafter Format-written, there is no significant additional effect of Format-written.
- When the readability properties are entered one at a time, there is a significant additional effect when Word length is entered as the second variable, but not when Information density is entered as the second variable.

Thus, to a large extent it is a common/overlapping effect of the three variables that creates a significant connection to demand of reading ability. However, the effect is primarily due to aspects of readability, in particular from the property Word length.

In addition, in the data we have used (from PISA 2003 and 2006) the different types of task properties are not independent. T-tests reveal that there are significant differences between tasks asking for a written answer and other tasks, regarding both the use of long words ($t(61)=2.127$, $p=0.037$) and also the noun-verb quotient ($t(55)=2.093$, $p=0.041$).

CONCLUSIONS

Regarding research question Q1, we have shown that two readability properties (Word length and Information density) and one task type property (Format-written) can in a statistically significant manner predict the demand of reading ability for a mathematics task. Thus, similarly as other studies (Bergqvist, 2009; Nortvedt, 2009; Roe & Taube, 2006), we also find a significant effect of the type of answer students should give on the demand of reading ability. However, regarding research question Q2, our results also show that there is primarily not a separate/genuine effect of the type of answer asked for but that the observed effect might be caused by a more general linguistic effect, since we observe the overlap in the effects from type of answer asked for and readability properties. Besides this common effect of different properties, there is a genuine effect of readability on the demand of reading ability, where the effect is mainly from the use of long words.

Thus, among the properties here analyzed it is primarily readability properties of a task that predict the task's demand of reading ability, in particular Word length and to some extent Information density (the noun-verb quotient). In order to reduce this, presumed unnecessary, demand of reading ability, you could try to use shorter words and avoid nominalizations. However, such 'easy fixes' can be pitfalls for several reasons. For example, the relationships between the studied variables are not necessarily causal. In addition, to isolate these types of changes can be difficult or even impossible, e.g. to replace a long word with a corresponding short one, without altering the meaning of the text.

Finally, regarding the effect of task format, since this property is not independent from the readability properties, perhaps the effect of task format is not due to the linguistic demand of writing an answer but solely caused by an indirect effect of readability. This hypothesis is congruent with our result showing no genuine effect of task format on the demand of reading ability. Since it seems possible to separate these types of task properties in the construction of tasks, there is a need for more analyses, either using another data set or selecting tasks to analyze from PISA, in order to more directly study a potential effect of task format on the demand of reading ability.

REFERENCES

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, Greece: PME.

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

Björnsson, C.-H. (1968). *Läsbarhet.* Stockholm, Sweden: Liber.

Einarsson, J. (1978). *Talad och skriven svenska: Sociolingvistiska studier.* Doctoral Dissertation. Lund: Ekstrand.

Gustafson-Capková, S., & Hartmann, B. (2006). *Manual of the Stockholm Umeå Corpus version 2.0.* Retrieved from http://www.ling.su.se/staff/sofia/suc/manual.pdf

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.

Liberg, C., & Forsbom, E. (2009). Text and language in assessment of mathematics and science. In A. Saxena & Å. Viberg (Eds.), *Multilingualism: Proceedings of the 23rd Scandinavian Conference of Linguistics, Uppsala University, 1-3 October 2008* (pp. 328–332). Retrieved from http://uu.diva-portal.org/smash/record.jsf?pid=diva2:275878

Lindberg, I., & Johansson Kokkinakis, S. (Eds.). (2007). *OrdiL: En korpusbaserad kartläggning av ordförrådet i läromedel för grundskolans senare år.* Gothenburg, Sweden: The Institute of Swedish as a Second Language, University of Gothenburg. Retrieved from http://hdl.handle.net/2077/20503

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

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.

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

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 from http://www.norden.org/pub/uddannelse/uddannelse/sk/TN2006523.pdf

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

Österholm, M., & Bergqvist, E. (in press). Methodological issues when studying the relationship between reading and solving mathematical tasks. *Nordic Studies in Mathematics Education.*