THE ROLE OF LINGUISTIC FEATURES WHEN READING AND SOLVING MATHEMATICS TASKS IN DIFFERENT LANGUAGES

Ewa Bergqvist[1], Frithjof Theens[1], and Magnus Österholm[1,2]

[1]Umeå Mathematics Education Research Centre (UMERC), Department of Science and Mathematics Education, Umeå University, Sweden

[2]Department of Mathematics and Science Education, Mid Sweden University, Sweden

ABSTRACT

The purpose of this study is to deepen the understanding of the relation between the language used in mathematics tasks and the difficulty in reading and solving the tasks. We examine issues of language both through linguistic features of tasks (word length, sentence length, task length, and information density) and through different natural languages used to formulate the tasks (English, German, and Swedish). Analyses of 83 PISA mathematics tasks reveal that tasks in German, when compared with English and Swedish, show stronger connections between the examined linguistic features of tasks and difficulty in reading and solving the tasks. We discuss if and how this result can be explained by general differences between the three languages.

1. INTRODUCTION

Tasks have a prominent role in mathematics education. They are a part of the teaching situation, used both for students to learn and for assessment of students' knowledge. Mathematics tasks are also used by agents external to the teaching situation, for assessment of students, teachers, schools, or countries, as in international assessments such as the Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS). Therefore, there is a strong need to use good tasks, whether used in learning situations or in assessments. Mathematics tasks have a focus on the learning or assessment of *mathematical* ability (i.e., acquired proficiency in mathematics). However, also other types of abilities could be needed to work with and solve a mathematics task. Reading ability is such an ability, especially for tasks in written form, which are very common in mathematics education. For example, it is often stated that tests intended to measure mathematical ability should not measure reading ability and that test constructors therefore should use simple wording (e.g., OECD, 2009, p. 116). Such a statement shows an attempt to separate reading ability from mathematical ability, which is problematic for two main reasons. Firstly, reading ability is always needed to solve a written task, since the students need to be able to read it to solve it. Secondly, and more importantly, mathematical communication is commonly considered to be one of the main aspects of mastering school mathematics, as can be seen in curriculum documents (e.g., NCTM, 2000; Niss & Højgaard, 2011) and in research frameworks (e.g., Lithner, Bergqvist, Bergqvist, & Boesen, 2010). That is, being able to read (and write) mathematics is seen as an important part of knowing mathematics. Still, it is important to avoid *unnecessary* demands of reading ability in mathematics tasks. Therefore, the relation between reading ability and mathematical ability is complex and important to examine in more detail, as is done in this study. In particular, we examine issues of *difficulty in reading*, which refers to a type of unnecessary demand of reading ability, together with issues of *difficulty in solving*, which refers to a holistic view of how difficult a task is to solve.

There are many different issues of language that are important within mathematics education. One is the existence of this relation between reading ability and mathematical ability. Another issue relates to the increasing globalization and internationalization of education and includes the use of many different languages in the same situations (Barwell, Barton, & Setati, 2007; Morgan, Craig, Schuette, & Wagner, 2014). For example, many students are second language learners and several languages can be used in the same teaching situation. Furthermore, mathematics tasks are translated to many different languages, both within countries and also in international comparative studies like PISA and TIMSS. In this study, we therefore focus on different languages by examining tasks that have been translated to many languages.

Research has shown that translated tasks can function statistically differently in different languages and therefore different language versions of a task might measure slightly different things (e.g., Ercikan & Koh, 2005). Empirical studies have shown that there can be many different reasons why different language versions of tasks measure different things. The reason can be pure translation errors, but sometimes it has been shown to be inherent properties of the languages that the tasks are formulated in (e.g., Allalouf, Hambleton, & Sireci, 1999). One type of inherent property concerns compound words. In some languages, many concepts are denoted with compound words, while in other languages concepts are denoted by using several separate words (cf. Pirkola, 2001), so that both word length and sentence length vary between languages. Another example of an inherent property is that subject-specific words might be more or less transparent in different languages. For example, the Chinese (Mandarin) word for *median* literally means *center number* (Han & Ginsburg, 2001), and similarly, the translation of *mean* is *medelvärde* in Swedish and *Mittelwert* in German, which both literally mean *middle value*.

Previous research described above highlights different issues concerning the relationship between language and mathematics, but also shows the need for more research on these issues.

2. Purpose and research questions

The purpose of this study is to deepen the understanding of the relation between *the language* used in mathematics tasks and the *difficulty in reading and solving* the tasks. By *language* we refer both to linguistic features of the tasks (e.g., wording and grammatical structure) and to the natural language used to formulate the tasks (e.g., English). This study therefore examines if there are any connections between different linguistic features on the one hand, and difficulty in reading and solving on the other hand, for mathematics tasks in English, German, and Swedish. In particular, we compare the results between the different languages. The overarching research question is:

> Do linguistic features of mathematics tasks relate to difficulty in reading and solving the tasks in different ways for tasks written in English, German, and Swedish?

The three languages were chosen for two reasons. Firstly, even though these languages are closely related, we know from earlier research that there are some important differences between them regarding several linguistic features (Sigurd, Eeg-Olofsson, & Van Weijer, 2004). For example, it is more common in both German and Swedish, compared to English, to create compound words (cf. Pirkola, 2001). Compound words make words longer in general, for example, *the bus station* is just one word in Swedish, *busstationen*, which combines the two words for bus and station, together with the suffix *-en* that corresponds to the determiner *the* in English. However, it is not clear whether longer or more words in a task might be related to some type of task difficulty, and therefore these are some of the issues studied in the present article. Secondly, the languages are chosen because of their closeness. We argue that large differences between tasks in these languages are less likely to be caused by bad translations and more likely to be caused by inherent and unavoidable properties of the languages. The reason is that it is easier to find more direct translations between such close languages, since they are very similar concerning both vocabulary (cf. Wichmann, Holman, & Brown, 2016) and structural properties (cf. Dryer & Haspelmath, 2013). We also argue that when the case of similar languages has been examined more closely, we are more equipped to focus on the more complicated cases, that is, comparing mathematics tasks in languages from different language families.

We focus on four different linguistic features of the tasks: word length, sentence length, task length, and information density. These features are chosen because they are quite often examined in connection to reading difficulties or complexity of texts. Another reason is that there are known differences between English, German, and Swedish, regarding at least the first two features (Sigurd et al., 2004).

The overarching question concerns whether these linguistic features are related to difficulty in reading and solving in different ways in these different languages, and therefore we formulate the following, more specific, research questions:

1. What linguistic features are only in some languages connected to *difficulty in reading* and *difficulty in solving*, respectively?
2. Is there variation between the three languages in how much the linguistic features explain the variation of *difficulty in reading* and *difficulty in solving*, respectively?

By *difficulty in reading* we refer to a measure of demand of reading ability (DRA) that addresses an unnecessary type of demand, which focuses on relations between reading ability and mathematical ability. By *difficulty in solving* we refer to a measure based on students' success rate of the task. The concepts of difficulty in reading and difficulty in solving are described more in the Background (Section 3.3). The concrete measures are then described in more detail in the Method (Section 4.2).

By answering these research questions, we gain knowledge about if the relations between linguistic features of mathematics tasks and task difficulty are different in different languages. This knowledge is central for exploring how properties of different languages can influence the mathematical experience for students. More specifically, through answering our research questions, we gain knowledge about how any differences between languages concern difficulties in the reading of the task or more general difficulties in the process of solving the task. This knowledge is central for exploring relationships between reading ability and mathematical ability, including issues of validity of assessment tasks, if students to a large extent need to rely on a general type of reading ability when solving mathematics tasks.

3. Background

This section addresses different issues relevant to the topic of the study, that is, the relation between *the language* used in mathematics tasks and the *difficulty in reading and solving* the tasks. By language we refer both to the natural language used to formulate the tasks (e.g., English), which is addressed in the first subsection, and to linguistic features of the tasks (e.g., wording and grammatical structure), which is addressed in the second subsection. In the third subsection, we address issues around difficulty in reading and solving mathematics tasks. In all three subsections, we discuss previous research and highlight aspects of theory and methodology that serve as a basis for the present study.

3.1 Multilanguage assessment

In the present study, we examine *multilanguage assessment* (sometimes denoted *multilingual* assessment), that is, assessment that is “administered in more than one language” (Ercikan, 2002, p. 199). In particular, we focus on the issues connected to comparing the results between student groups taking different language versions of a test. Comparing test results between groups is generally complicated, since many different factors can be involved in students’ performances, besides their knowledge of the particular subject. For example, the students’ gender and ethnicity, as well as their curricular, cultural, and language background can influence their results (e.g., Roth, Ercikan, Simon, & Fola, 2015). The issue of comparing task results across languages is even more complex. One reason is that it often includes comparing between countries and cultures (Harkness et al., 2010). In addition, the comparability of tasks translated to different natural languages is related to several other issues particular for multilanguage assessment. Firstly, the quality of the translation is important, since low quality translations might “cause problems in comparability and equivalence” (Ercikan, 2002, p. 199). Secondly, even high quality translations might function differently due to inherent properties of the natural languages involved, and the possibility to create high quality translations is connected to how similar the natural languages are. Natural languages can differ in many aspects, for example, regarding vocabulary and grammatical structures. There are also less obvious differences, for example, that a particular concept can be transparent in one language but more obscure in another (Leung, 2014). For example, the literal translation of the

Chinese word for *quadrilateral* is *four-side-shape* (Han & Ginsburg, 2001), and similarly, the Swedish translation *fyrhörning* and the German translation *Viereck* are both built from the words *four* and *corner* (or *vertex*). Thus, if the question 'How many sides does a quadrilateral have?' is translated word-by-word to Chinese, the question becomes much easier, since the Chinese formulation would literally correspond to 'How many sides does a four-sided-shape have?' Another type of difference between natural languages concerns orthographical depth, that is, how transparent the letter-to-sound mapping of the language is. For example, in Finnish it is easy to correctly pronounce a word that you have never read before, but in English it is often difficult or even impossible (Ziegler et al., 2010). Difference in orthography is a central reason for differences in reading acquisition between languages and has been studied extensively in relation to reading acquisition (Ziegler & Goswami, 2005). This difference could affect the transparency of translated texts, so that the text is more transparent in one of the languages. In conclusion, multilanguage assessments struggle with many different issues of comparability between student groups.

The *type* of test is also necessary to take into consideration during multilanguage assessment (Ercikan, 2002). For example, psychological tests often rely only on everyday language while tests in content areas also include and focus on different linguistic registers (see Halliday, 1975). Achievement tests in mathematics use the particular mathematical register, including technical vocabulary, multiple semiotic systems, and certain grammatical patterns (Schleppegrell, 2007). In addition, mathematics tasks (in particular word problems) have their own particularities and can be seen as a linguistic genre (Gerofsky, 1999). Different linguistic genres or registers might demand different types of literacy. For example, it is often stated that reading mathematics demands a specific type of reading ability (Burton & Morgan, 2000; Cowen, 1991; Fuentes, 1998; Konior, 1993; Shanahan & Shanahan, 2008). In addition, McKenna and Robinson (1990) define the concept of content literacy as consisting of three components: general literacy skills, content-specific literacy skills, and prior knowledge of content. Similar divisions are made by, for example, Behrman and Street (2005). Thus, another dimension of complexity is present in multilanguage assessment when the assessment focuses on content areas.

In addition, the mathematical register differs between different natural languages and some of these differences might affect students' experience and comprehension of mathematics and mathematical texts, such as during multilanguage assessment. As mentioned earlier, a word can be more transparent for the students in one language than in another, and this type of difference also concerns mathematics-specific words (see previous examples). Also, the written numbers are represented with arabic numerals using positional notation in the same way in many (western) languages, but still, the structure of how they are pronounced is very different. For example, the number 32 is pronounced *zweiunddreißig* in German, corresponding to *two-and-thirty*, starting with the unit digit, but in English and in Swedish, pronunciation starts with the tens digit. This difference has been the focus of many research studies, which shows that this naming in the number system affects students' processing of numbers, especially for younger children (e.g., Nuerk, Weger, & Willmes, 2005; Pixner, Moeller, Hermanova, Nuerk, & Kaufmann, 2011). Another difference is that the number words between 11 and 19 are pronounced in a regular way in some languages and in an irregular way in other, where the decomposition in tens and units is hidden (Fuson & Kwon, 1992; Geary, Bow-Thomas, Liu, & Siegler, 1996). For example, in several Asian languages the number 13 is pronounced exactly as *ten-three* and 11 as *ten-one*, marking that it consists of a ten and a three or a one. In English there is not the same obvious correspondence with *thirteen* and *eleven*. The same issue also exists for higher numbers, where the decomposition can be more or less hidden (cf. Liu, Lin, & Zhang, 2016). Empirical studies have shown that these types of differences seem to affect the type of calculation strategies used by students and also the speed of number processing, at least for younger students (Geary et al., 1996). It has also been shown that an awareness of the morphology of number words is connected to the ability to make calculations for younger children (Liu et al., 2016).

The issues addressed above highlight the complexity of multilanguage assessment. Statistical methods are sometimes used to examine this complexity, and differential item functioning (DIF) analysis is a standard method for examining statistical differences between test tasks (items) for different groups of students (Zumbo, 1999). Groups are often divided based on gender, performance, or language proficiency (e.g., Heppt, Haag, Böhme, & Stanat, 2015), but DIF analysis can also be used to compare groups that have taken different language versions of a task, that is, for multilanguage assessment. From a theoretical perspective, a task is said to show DIF if the probability to answer a task correctly is not the same for members with equal ability, but from different groups, which "means that there is some sort of systematic but construct irrelevant variance that is being tapped by the test or measure" (Zumbo, 1999, p. 34). DIF studies can show if, but not why, tasks in a multilanguage assessment function differently in different languages. One possible reason for DIF in multilanguage assessment is that the translation is bad. Another possible reason is that there are inherent properties of the languages that affect how the mathematics can (or must) be presented, as described above, in which case it is likely to be impossible to completely erase the DIF. For long, there was little research focusing on the causes of DIF in multilanguage assessment (e.g., Allalouf et al., 1999; Gierl & Khaliq, 2001). More recently, there are studies examining the possibility to complement DIF analyses with, for example, think aloud protocols (Ercikan, Arim, Law, Domene, Gagnon, & Lacroix, 2010; Roth, Oliveri, Sandilands, Lyons-Thomas, & Ercikan, 2013) or bilingual experts reviewing the items (Allalouf et al., 1999; Ercikan, Gierl, McCreith, Puhan, & Koh, 2004). There are also other types of statistical methods used to examine the relation between performance on mathematics tasks and language demands, for example, Bailey (2005) studies the correlation between linguistic demands of mathematics tasks and the difference in performance by English Language Learners (ELLs) and non English Language Learners (non-ELLs). These different studies, whether using DIF or other methods, tend to focus on different *types* of reasons, using broad categories, such as linguistic reasons and curriculum reasons. They also tend to be more exploratory concerning which types that are more relevant in multilingual assessment. We have not found any studies on multilanguage assessment that choose certain features of tasks, and then analyse them more in-depth, concerning what role these features have in different language versions and concerning relations to reading and solving the tasks.

3.2 Linguistic features

In this study we focus on four different linguistic features of the tasks: word length, sentence length, task length, and information density. Below we present previous research regarding these features and how we measure them. In Section 4.3, we describe further the reasons for this choice, including references to previous studies using the same or similar measures, and the practical issues connected to the measurements. In addition, we present as an example, all calculations of all variables for an excerpt of a PISA task in Table 1.

Much of the previous research described below uses empirical data in English or does not mention the language at all (e.g., Helwig, Rozek-Tedesco, Tindal, Heath, & Almond, 1999). Results from such studies cannot necessarily be generalized to other languages, for example, due to differences between languages concerning the specific features that are analysed. In our study, we address this issue by analysing three different languages concerning the same relations between linguistic features and difficulty.

3.2.1 Word length

The variable *word length* is often one of several aspects included in frameworks of linguistic complexity (e.g., Abedi, Leon, Wolf, & Farnsworth, 2008) and in readability formulae (e.g., Lenzner, 2014). Studies of memory have reliably shown that it is easier to remember lists of short words than lists of long words, although it is not completely clear what causes this so called *word length effect* (Jalbert, Neath, Bireta, & Surprenant, 2011). The presence of long words in a text has for a long time been seen as something that creates difficulties for readers (e.g., see Flesch, 1948; Lenzner, 2014). However, word length, together with some other features, might be indices for linguistic complexity, rather than (necessarily) direct measures of the actual difficulty (Abedi, Hofstetter, Baker, & Lord, 2001). For example, long words are more often uncommon and more morphologically complex than short words, so the reason that word length in some cases correlates with reading difficulty could sometimes be explained by the existence of uncommon or morphologically complex words. Furthermore, the relation between word length and reading comprehension can vary between groups, for example, between weak and strong readers (Marmurek, 1988).

Word length in relation to the solving of mathematics tasks has been examined empirically in many different ways. Some studies calculate the mean length (in letters) of the words in mathematics tasks (e.g., Lepik, 1990; Lin, 2012) while others count the number or percentage of long words, that is, words longer than a particular number of letters or syllables, in the tasks (e.g., Helwig et al., 1999; Lepik, 1990; Norgaard, 2005; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006; Österholm & Bergqvist, 2012b). Several studies have shown no statistically significant correlations between different measures of word length and different types of difficulty (e.g., solution frequency) in mathematics tasks. However, Lin (2012) found that the mean word length (in letters) for mathematics tasks in English as well as the number of words with only one syllable (i.e., short words) correlated with the solution frequency for all student groups (some with learning disabilities). In addition, also for mathematics tasks, but in Swedish, Österholm and Bergqvist (2012b) found that the percentage of words with more than six letters significantly correlated with an unnecessary demand of reading ability (DRA, see Section 4.2).

In this study, we calculate four different variables in relation to *word length* for a task: *fraction of words longer than six letters, average word length in letters, fraction of multisyllabic words (i.e., words longer than one syllable),* and *average word length in syllables.* We use several different variables since previous studies have used a variety of measures and there has been a mixture of results in such studies.

3.2.2 Sentence length

Besides word length, sentence length is also used in many readability formulae. It is often measured as the average number of words per sentence (Glazer, 1974; Haag, Heppt, Roppelt, & Stanat, 2014), but sometimes also as letters or literals per sentence (Lepik, 1990) or the number of units (not only words but also abbreviations and formulae) per sentence (Kulm, 1971). The focus on units, and not only ordinary words, is relevant in particular for mathematics texts, where many abbreviations or symbols can be part of a sentence. As with word length, sentence length can be a difficulty in itself, but it can also be seen as an indicator of another type of difficulty. In particular, since longer sentences often have a higher syntactic complexity, for example, by using subordinate clauses, sentence length is an easily accessible indicator of the syntactic difficulty of a text (Glazer, 1974; Lenzner, 2014). In earlier research, sentence length showed no significant correlation with an unnecessary demand of reading ability for mathematics PISA tasks in Swedish (Österholm & Bergqvist, 2012b). Nevertheless, because of differences between languages regarding sentence length (Sigurd et al., 2004), it is worth to investigate the effects of sentence length for different languages.

In this study, *sentence length* for a task is measured as *the average number of words, but also other units* such as abbreviations, numbers, and formulae, *per sentence* in the task (Kulm, 1971).

3.2.3 Task length

Task length has in previous research been measured in many different ways, such as the number of words, the number of unique content words (i.e., nouns, verbs etc.), and the number of sentences (Abedi et al., 2008). Earlier research has on some occasions shown that there is a relation between the length of mathematics tasks and student performance, in Swedish (Bergqvist, Dyrvold, & Österholm, 2012) and in English (Wolf & Leon, 2009), while other studies of tasks in English have not shown any significant correlation between these aspects (e.g., Lepik, 1990; Shaftel et al., 2006). Jerman (1974) designed mathematics tasks in English of different length but with other factors, such as order of operations and task structure, kept constant. The study showed that in some cases, but not all cases, task length to some extent explained task difficulty. The conclusion was that it was not the task length per se that made some tasks more difficult, but somehow task length in relation to other factors. Task length, in the same way as word length and sentence length, might serve as an indicator of other types of complexity that can make the tasks more difficult and is therefore worth to examine in relation to difficulty. In addition, for language learners, longer mathematics tasks seem to be more difficult both in English (Abedi et al., 2008; Abedi, Lord, & Plummer, 1995) and in German (Haag, Heppt, Stanat, Kuhl, & Pant, 2013). Also, French versions of PISA reading tasks, that are longer in French than in English, are a little more difficult than the English versions (Grisay, 2003).

We measure *task length* as *the total number of words in the task* (see White, 2012).

3.2.4 Information density

Intuitively it seems reasonable that the density of information can make a text more difficult to understand due to the need to unpack the information. At the same time, it can be seen as necessary to pack information in smaller units, to be able to address more complex issues. For example, to condense an activity or a process into a single noun (i.e., an act of objectification or reification) can be seen as necessary in the learning of mathematics (Sfard, 1991). That is, it is relevant to focus on information density in relation to difficulty, including the potential connections to a relevant (necessary) type of difficulty. However, there is no common way to measure the level of information density in a text. But several different types of measures have shown to correlate with comprehension difficulty, for example, when measuring the number of distinct concepts per sentence (Best, Ozuru, & McNamara, 2004), when including or excluding background details (Botta, Pingree, & Hawkins, 1993), when measuring the proportion of lexical, as opposed to grammatical, words (Perfetti, 1969), and when measuring the percentage of sentences relevant for an instructional objective (Rothkopf & Kaplan, 1972). Specifically for mathematics tasks, the noun-verb quotient, as a measure of information density (see Einarsson, 1978) has been shown to correlate with an unnecessary demand of reading ability for mathematics PISA tasks in Swedish (Österholm & Bergqvist, 2012b). This measure, unlike some of the other measures of information density, can vary between translated task versions, since it is a measure at lexical level, and it is therefore relevant to study in multilanguage assessment.

Information density is in this study measured as the noun-verb-quotient, that is, the *ratio between the number of nouns and the number of verbs* in the sentences of a task (Einarsson, 1978).

3.3 Difficulties in reading and solving mathematics tasks

International frameworks and curriculum documents describing knowledge in school mathematics include aspects of communication as part of mathematical competence (e.g., Lithner et al., 2010; NCTM, 2000; Niss & Højgaard, 2011). Based on this communicative aspect of mathematics, some of the demands of reading ability that a mathematics task puts on students are both reasonable and *necessary*. For example, the students should be able to read long words included in the mathematical vocabulary (such as *equation*) and also to interpret grammatical constructions representing mathematical relations (such as *twice the size of*), even if they are more unusual in colloquial language. Still, there can also be *unnecessary* demands, for example, if the solving of a task demands that the student understands very difficult words that are not included in the mathematics register. Separating between necessary and unnecessary demands, both theoretically and methodologically, is important since mixing them up could lead to incorrect conclusions regarding the quality of mathematics tasks, and thereby producing assessments of low quality and incorrect information about students' knowledge.

We use a simple theoretical model that describes students' reading ability and mathematical ability as two, partially overlapping, types of abilities. In Figure 1, area B illustrates the part of reading ability that is relevant, and potentially specific, for mathematics. Area C illustrates the part of reading ability that is not part of mathematical ability. If a student must use the ability in area C to solve a mathematics task, this can be seen as a sign of low validity of the task since this type of ability is not part of mathematical ability. Each arrow in Figure 1 symbolizes how much of the variation of the success rate of a task can be explained by a certain type of ability (area A, B or C). Necessary reading demands in a mathematics task then correspond to the arrow from area B, while unnecessary reading demands correspond to the arrow from area C. Based on this model, we use the success rate of a mathematics task as a measure of the task's *difficulty in solving* and the size of unnecessary reading demands (arrow from area C) as a measure of the task's *difficulty in reading*. This is of course a very simple model, since the relationship between mathematical ability and reading ability is complex. For example, empirical studies show that "the relationship is stronger for low math ability students and weaker for high math ability students" (Chen & Chalhoub-Deville, 2015, p. 596). Still, since we use the model at group level it can be informative regarding large-scale assessment.

The success rate is used as a measure of a holistic perspective on the level of difficulty when solving a task. This measure includes any type of difficulty encountered by students, regardless of which type of ability is needed to solve the task. If the presence of a certain feature of a task correlates with this measure of difficulty, we cannot conclude what type of difficulty this feature seems to create, but only that the feature seems to make tasks more difficult. In particular, we cannot decide only from such an analysis whether this feature is necessary or unnecessary to include in mathematics tasks. On the other hand, by focusing on unnecessary reading demands (arrow from area C), we can produce empirical results that give more direct information about features that are, at least potentially, unnecessary to include in mathematics tasks. That is, if the presence of a certain feature of a task correlates with our measure of difficulty in reading, which focuses on the use of an unnecessary type of reading ability, we can conclude that this feature seems to cause lower validity in mathematics tasks.

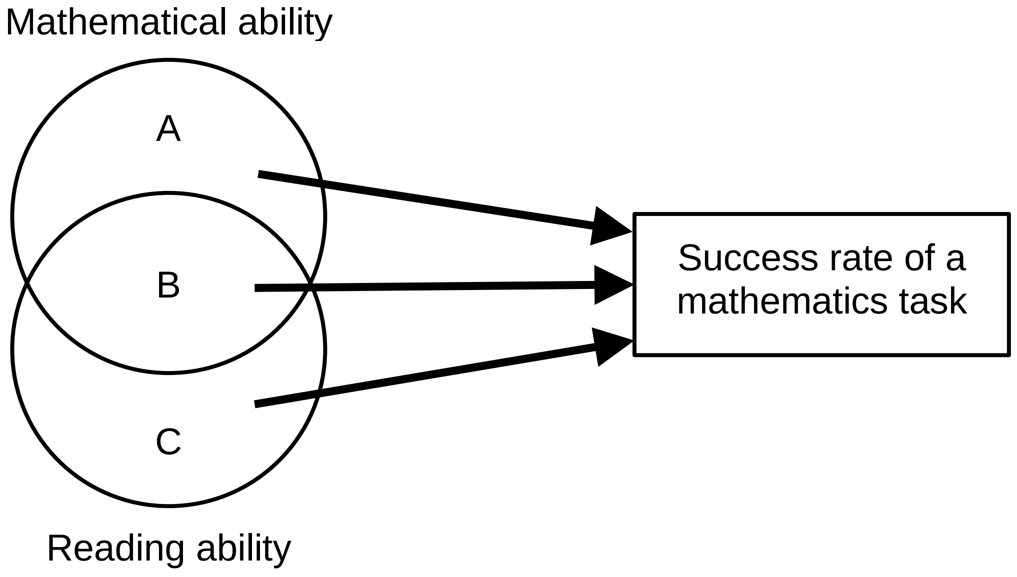

Figure 1. Schematic illustration of a theoretical model describing relations between abilities and the interpretation of necessary reading demands (arrow from area B) and unnecessary reading demands (arrow from area C).

Many statistical methods that are used for examining the relations between linguistic features of tasks and the difficulties of the tasks cannot separate between necessary and unnecessary reading demands. One type of method is the use of correlations between the existence of certain linguistic features of tasks and the students' results, to identify linguistically problematic tasks (e.g., see Roe & Taube, 2006). The problem is not the type of method per se, but the interpretation of the results and the conclusions that sometimes are drawn. A concrete (but simplified) example is when the existence of long words correlates with task difficulty, and the conclusion is that the tasks are difficult because the words are long and that the tasks therefore have an *unnecessary* demand of reading ability. The reason that the tasks are difficult could instead be that cognitively more advanced mathematical words often are longer, in which case the tasks are more difficult not because of the long words, but because of difficult mathematical vocabulary, that is, a *necessary* demand of reading ability. Thus, this method cannot separate between necessary and unnecessary reading demands. However, there are other methods that in theory can do this separation, but cannot do it in a reliable manner, as shown in a comparison between different methods (Österholm & Bergqvist, 2012a). This comparison of methods also shows that it is possible to use a principal component analysis (PCA) to separate the unnecessary demands from the necessary, in a valid and reliable manner. In this study we therefore use this method, by which it is possible to measure the influence of reading ability when the influence of mathematical ability has been excluded. This analysis results in the quantitative measure that we label as *demand of reading ability* (DRA), which is described in more detail in Section 4.2.

4. METHOD

To investigate the relationship between linguistic features of mathematics tasks and the difficulty in reading and solving the tasks in different languages, we performed an analysis in three steps. Firstly, we calculated measures of difficulty in reading and difficulty in solving each task in each language version. Secondly, we calculated the values for the measures of the linguistic features of each task in the different language versions. Thirdly, we performed statistical analyses, using correlations and regressions, to answer the research questions. Below, we present the selected data and then describe each of the steps of the analysis in more detail.

4.1 DATA SELECTION

The mathematics tasks used in this study are the tasks of the 2012 PISA assessment, used in USA, Germany, and Sweden. Using PISA tasks has several advantages.

Firstly, the translation of these tasks was made by professionals according to a rigorous procedure specified by the OECD (2010). In particular, translations are made from one English source version and also from one French source version, and these translations are then reconciled. The tasks used in USA are only adapted from the English source version. Through these procedures, pure translation errors that might affect the results are avoided as far as possible.

Secondly, a large number of students worked with the tasks; there were 4978 participating students in the USA, 5001 in Germany, and 4736 in Sweden. Every student worked with a booklet containing about 30 % of the tasks, which means that there are about 1500 student results available for each task in each language.

Thirdly, and most importantly, the 2012 PISA tasks were used in this study because besides the mathematics tasks there were also 44 reading tasks in the assessment. This combination of mathematics tasks and reading tasks, solved by the same students, makes it possible to calculate, for each mathematics task, a measure of difficulty in reading through the demand of reading ability (DRA), as described in the next section.

Of the 84 mathematics tasks used in PISA 2012, we had to exclude one task from the analysis due to too few answers from students in Sweden. One of the 44 reading tasks was deleted in the German version according to the Technical Report (OECD, 2014). Therefore we excluded this item for all countries and accordingly used 43 reading tasks in the analyses. As explained below, we also excluded all mathematics tasks with a negative DRA measure when the statistical analysis included DRA. In total, we included 62 mathematics tasks in English, 63 in German, and 60 in Swedish in the analysis for DRA. For difficulty in solving, 83 mathematics tasks were used in the analyses.

4.2 Measurements of difficulty in reading and solving mathematics tasks

The measure of the difficulty in solving a mathematics task was calculated as one minus the success rate.

Demand of reading ability was used as a measure of the difficulty in reading a mathematics task. The idea with demand of reading ability is to focus on the part of reading ability that is not part of mathematical ability, that is, the *unnecessary* reading demands (see area C in Figure 1). For each language, we created a matrix consisting of the students' scores, with all mathematics and reading tasks as columns and all individual students as rows. As mathematics tasks mainly measure mathematical ability and reading tasks mainly measure reading ability, one can expect correlations mostly *within* the results of the mathematics tasks on the one hand and mostly *within* the results of the reading tasks on the other hand. However, to some degree, the results from each mathematics task will correlate also with the results from the reading tasks. The matrix for each language was therefore analysed to show how strongly the results for each mathematics task correlate with the results from the group of reading tasks. The analysis creates a quantitative measure of this correlation for each mathematics task and we call this measure demand of reading ability (DRA).

Describing the creation of the measure of DRA in more technical terms, we did the following. We performed a Principal Component Analysis (PCA), separately for each of the three languages, on the matrix consisting of the students' scores on both mathematics and reading tasks. It has been shown that the use of a PCA is the best method to achieve high validity and reliability in this type of analyses of relations between mathematical ability and reading ability (Österholm & Bergqvist, 2012a). We here address the main steps and arguments for the use of a PCA in this type of analysis. A PCA is suitable in analyses of different dimensions in data, which in our case refer to two different abilities, in mathematics and reading. Therefore, we expected the mathematics tasks to load mainly on one component (corresponding to mathematical ability) and the reading tasks on another component (corresponding to reading ability) and therefore we extracted the first two components. We used an oblique rotation (Promax) in the PCA since mathematical ability and reading ability are not independent. Calculations of DRA showed that the majority of the mathematics tasks (51–58 % for the three languages) have a high loading only on the first component and a majority of the reading tasks (60–67 % for the three languages) have a high loading only on the second component. This result supports our expectation that we have two main components that correspond to mathematical ability and reading ability. For each mathematics task, the loading value on the component corresponding to reading ability was then interpreted as a measure of the task's DRA. We used the loading values from the pattern matrix since these ”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, 2007, p. 627). Therefore, 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, which corresponds to the arrow from area C in Figure 1. That is, the loading value gives a quantitative measure of the unnecessary reading demands for each mathematics task. Finally, since this study focuses on linguistic features connected to *demand* of reading ability we only included tasks with a positive loading on the component corresponding to reading ability in the statistical analyses that include DRA.

4.3 Linguistic features

The linguistic features investigated in this study are word length, sentence length, task length, and information density. We analysed these linguistic features in mathematics tasks formulated in English, German, and Swedish. The analyses were mainly carried out by the second author, who is a native German speaker and fluent in both English and Swedish. The first and third authors are native Swedish speakers and fluent in English, and supported the third author's analyses in Swedish. The authors were also supported by an Advisory Board of experts on language in education, including native speakers of English, German, and Swedish.

Generally, when a task is translated to another language, figures (like tables and formulas) and images are kept unchanged, except for some occasional words that are translated. Our starting point is therefore that we examine, first and foremost, the natural language part of tasks in different languages, because this is mainly what is changed in the translation. More concretely, examining the natural language means that we compare complete words written with letters that are encoding sounds (i.e., built up by phonemes). That is, abbreviations (e.g., ATM, CO_2, ITU), abbreviated units (e.g., cm, km/h, dl), numbers written in numerals, and other symbols (e.g., %) did not count as words. However, as described and motivated below, we made an exception concerning sentence length. For the variables of word length and task length, we included all words of the task, for example, also words in tables, pictures, and graphs. But for sentence length and information density we only included complete sentences, since these variables address features at this linguistic level. More details regarding how variables were calculated for each linguistic feature are given in the following subsections. An example from PISA, where these variables are calculated, is given in Table 1. In total, seven different variables were defined and calculated.

Many of the PISA tasks start with an introductory text, shared by a number of subtasks, which can vary in length from only a few words up to several sentences. When we calculated the measures of the different linguistic features for each subtask, the text of the introduction was included in each subtask, since it can be necessary to read both the introduction and the text of the subtask to be able to solve it.

4.3.1 Word length

In this study, we calculated four different variables for measuring different aspects of word length: *fraction of words longer than six letters*, *average word length in letters*, *fraction of multisyllabic words*, and *average word length in syllables.* We counted both syllables and letters since these measures are connected to different issues of decoding (length of sound and physical length, respectively) that might make long words difficult to read. Also, we counted both the *average word length* and the *fraction of long words* for each task, since these aspects measure slightly different things. We chose not to count the *total number* of long words, since tasks with many words probably contain a higher number of long words than shorter tasks (but not necessarily a higher *fraction* of long words), and therefore the total number of long words probably correlates with *task length*, which is measured separately (see below). Regarding *fraction of words longer than six letters*, we chose this measure since it was used in an earlier study (Österholm & Bergqvist, 2012b) where it correlated significantly with demand of reading ability for Swedish PISA tasks in mathematics. Regarding the measure of long words as *multisyllabic* words, it was chosen since it was used in previous studies (e.g., Helwig et al., 1999).

4.3.2 Sentence length

To measure sentence length, we examined only complete sentences (usually with a subject and a predicate). Sentence length is often used as an indicator of syntactic difficulty, and to create a measure that captures the complete sentence, we counted not only ordinary words, but also other units such as abbreviations, numbers, and formulae. Sentence length was measured as the *average number of* such *units per sentence*, which has been used in previous research, in particular for mathematics texts (Kulm, 1971).

4.3.3 Task length

We see task length primarily as an issue of amount of content in the task, that is, it is potentially more difficult to handle all the content of a long task than a short task. Therefore, we measured task length as the *total number of words in the task* (and not letters or syllables), which is a measure also used in previous studies (e.g., White, 2012). This measure is not suitable for all types of analyses, for example, when comparing different versions of a task in the same language, where the same amount of content could be described using different number of words. However, since we use the measure to compare different tasks, we see this measure as suitable.

4.3.4 Information density

Information density was calculated as the *noun-verb-quotient* (*nominalkvot* in Swedish), that is, the ratio between the number of nouns and the number of verbs in the sentences of a task (Einarsson, 1978). We used only complete sentences when calculating information density. Tables, figures and diagrams often contain long lists of nouns and phrases without verbs, thus these texts were excluded from the analysis, unless they were complete sentences. We focused on if the words functioned as nouns or verbs in the sentence, for example, a participle that functioned as adverb did not count as a verb.

4.4 Statistical analyses

Research question 1 concerns connections between variables and therefore we used correlations to answer it. For each of the three languages, we calculated correlation coefficients between each variable measuring a linguistic feature and difficulty in reading (i.e., DRA) and also between each variable measuring a linguistic feature and difficulty in solving (i.e., one minus success rate). Among all correlations, we focused on those that were statistically significant (at level 0.05), and compared these between the languages, to see which correlations are significant in only some of the languages. Research question 2 concerns the amount of explained variance and therefore we used regressions to answer it. For each language, and separately for difficulty in reading and difficulty in solving used as dependent variable, we calculated a regression model where we inserted all seven variables measuring linguistic features as independent variables. We then compared the size of total explained variance between the three natural languages. In the statistical analyses including DRA we only used the tasks with a positive loading on the reading component since tasks with negative loading do not have a *demand* of reading ability. This resulted in 62 mathematics tasks in English, 63 in German and 60 in Swedish included in the analysis for DRA. For difficulty in solving, all 83 mathematics tasks were used in the analyses.

Table 1. Example from PISA (part of task PM903Q03) where all variables have been calculated based only on this excerpt.

English: An infusion with a drip rate of 50 drops per minute has to be given to a patient for 3 hours. For this infusion the drop factor is 25 drops per milliliter. What is the volume in mL of the intravenous drip?
German: Eine Infusion mit einer Tropfrate von 50 Tropfen pro Minute muss einem Patienten 3 Stunden lang verabreicht werden. Für diese Infusion ist der Tropffaktor 25 Tropfen pro Milliliter. Wie groß ist das Volumen der Infusion in ml?

<table>
<tr><td colspan="4"><b>Swedish:</b>
En infusion med en dropphastighet på 50 droppar per minut måste ges till en patient under 3 timmar. För den här infusionen är droppfaktorn 25 droppar per milliliter.
Vad har infusionen för volym i ml?</td></tr>
<tr><td><b>Variable</b></td><td><b>English</b></td><td><b>German</b></td><td><b>Swedish</b></td></tr>
<tr><td>Word length (letters/word)[a]</td><td>4.1 (154 letters in 38 words)</td><td>5.5 (180 letters in 33 words)</td><td>5.0 (156 letters in 31 words)</td></tr>
<tr><td>Word length (syllables/word)[a]</td><td>1.4 (54 syllables in 38 words)</td><td>1.9 (63 syllables in 33 words)</td><td>1.8 (56 syllables in 31 words)</td></tr>
<tr><td>Word length (fraction of words with letters&gt;6)[a]</td><td>0.13 (5/38)</td><td>0.36 (12/33)</td><td>0.29 (9/31)</td></tr>
<tr><td>Word length (fraction of words with syllables&gt;1)[a]</td><td>0.26 (10/38)</td><td>0.55 (18/33)</td><td>0.45 (14/31)</td></tr>
<tr><td>Sentence length
(words/sentence)[b]</td><td>14.0 (42 words in 3 sentences)</td><td>12.3 (37 words in 3 sentences)</td><td>11.7 (35 words in 3 sentences)</td></tr>
<tr><td>Task length
(words/task)[a]</td><td>38</td><td>33</td><td>31</td></tr>
<tr><td>Information density
(nouns/verbs)</td><td>3.75 (15/4)</td><td>3.25 (13/4)</td><td>3.25 (13/4)</td></tr>
</table>

[a] For the calculations of word length and task length, only complete words written with letters that are encoding sounds were included, that is, the numbers 50, 3, and 25 and the abbreviation mL were excluded (see section 4.3).

[b] For the calculation of sentence length, all words and other units were included, also numbers and abbreviations (see section 4.3).

5. Results

This section has three parts. Firstly, we present some descriptive statistics of the data and the measures used. Secondly, we answer research question 1 by presenting statistics focusing on connections between measures of linguistic features and the difficulty in reading (i.e., DRA) and difficulty in solving (i.e., one minus success rate). Thirdly, we answer research question

2, regarding how much of the total variance of difficulty in reading and difficulty in solving is explained by linguistic features. In section 6, we discuss and try to explain the findings.

5.1 Descriptive statistics

Table 2 gives an overview of the variables used in this study. As expected, words tended to be longer and sentences tended to be shorter in German and Swedish compared with English. These patterns are also visible in the example given in Table 1, and the example highlights different reasons for these patterns. In particular, on two occasions, the creation of compound words is evident in German and Swedish. *Drip rate* is translated into *Tropfrate* in German and into *dropphastighet* in Swedish, and *the drop factor* is translated into *der Tropffaktor* in German and into *droppfaktorn* in Swedish. Furthermore, Table 2 shows that the average levels of difficulty in reading and difficulty in solving are very similar in the three languages.

Table 2. Mean and standard deviation (SD) of the variables measured for the mathematics tasks.

	English		**German**		**Swedish**	
Linguistic feature	**mean**	**SD**	**mean**	**SD**	**mean**	**SD**
Word length (letters/word)	4.63	0.42	6.03	0.62	5.36	0.52
Word length (syllables/word)	1.50	0.17	2.01	0.23	1.96	0.23
Word length (fraction of words with letters>6)	0.20	0.06	0.34	0.07	0.28	0.07
Word length (fraction of words with syllables>1)	0.33	0.10	0.55	0.06	0.52	0.08
Sentence length (words/sentence)	14.62	3.95	13.63	3.62	13.15	3.70
Task length (words/task)	99.37	48.76	90.33	45.02	87.81	45.67
Information density (nouns/verbs)	2.89	1.25	2.45	1.00	2.36	0.94

Difficulty in reading (DRA)	0.16	0.21	0.16	0.21	0.14	0.20
Difficulty in solving (one minus success rate)	0.56	0.24	0.49	0.23	0.55	0.24

5.2 Research question 1

The first part of research question 1 concerns which linguistic features are connected to *difficulty in reading* only in some languages. Table 3 shows that there are statistically significant correlations between linguistic features and difficulty in reading for tasks in German, while the correlations for tasks in English and Swedish are not close to being statistically significant (p>0.3 for all correlations for English and Swedish). The connections to difficulty in reading for German tasks exist for different measures of word length (in particular when using a cut-off value for long words), and for information density. The correlation between information density and difficulty in reading is positive, which means that an increase in information density is connected to an increase in demand of reading ability. The connection between word length and difficulty in reading is in the reverse direction, that is, an increase in word length is connected to a decrease in demand of reading ability.

Table 3. Correlation coefficients between variables measuring linguistic features and difficulty in reading (demand of reading ability, DRA) for three language versions of mathematics tasks, only including tasks with positive DRA, since these indeed have a demand of reading ability. Statistically significant correlations are marked with * (p<0.05) or ** (p<0.01).

Linguistic feature	**English** **(N = 62)**	**German** **(N = 63)**	**Swedish** **(N = 60)**
Word length (letters/word)	0.057	-0.263*	0.003
Word length (syllables/word)	0.045	-0.180	-0.027
Word length (fraction of words with letters>6)	0.117	-0.358**	-0.128
Word length (fraction of words with syllables>1)	0.032	-0.333**	-0.128

Sentence length (words/sentence)	-0.043	0.076	0.136
Task length (words/task)	-0.004	0.094	-0.015
Information density (nouns/verbs)	-0.018	0.270*	0.013

Research question 1 also concerns which linguistic features are connected to *difficulty in solving* only in some languages. The results show statistically significant correlations between information density and difficulty in solving for tasks in German and Swedish, but not for tasks in English (see Table 4). However, for tasks in English, the correlation between information density and difficulty in solving is close to statistically significant (p=0.09). All these correlations are negative, that is, an increase in information density is connected to an increase in success rate (a decrease in difficulty in solving). One other correlation is very close to being statistically significant: one measure of word length (syllables>1) for German tasks (p=0.052). This correlation is positive, that is, an increase in word length is connected to an increase in difficulty in solving.

Table 4. Correlation coefficients between variables measuring linguistic features and difficulty in solving (one minus success rate) for three language versions of 83 mathematics tasks. Statistically significant correlations are marked with * (p<0.05) or ** (p<0.01).

Linguistic feature	**English**	**German**	**Swedish**
Word length (letters/word)	-0.027	0.059	-0.041
Word length (syllables/word)	-0.013	0.094	-0.068
Word length (fraction of words with letters>6)	-0.101	-0.031	0.093
Word length (fraction of words with syllables>1)	-0.014	0.214	-0.035

Sentence length (words/sentence)	0.065	0.040	0.075
Task length (words/task)	0.155	0.154	0.170
Information density (nouns/verbs)	-0.186	-0.288**	-0.264*

A correlation is a single number that quantifies the strength of the relation between two variables. In this study, we focus on correlations that are statistically significant at level 0.05, which means that there is always a 5 % risk that the correlation coefficient is significant even if there is no actual relationship between the involved variables. In a study like this, where many different correlations are calculated, there is therefore always a risk that some correlation coefficients appear significant by coincidence. However, in this study, the observed significant correlations do not occur randomly among the 42 correlations investigated, but are concentrated to German, regarding difficulty in reading, and to information density, regarding difficulty in solving. This indicates that the correlations most likely show real effects and are not significant by coincidence.

5.3 Research question 2

Research question 2 asks if there are differences between the three languages in how much of the variation of difficulty in reading and difficulty in solving, respectively, is explained by the linguistic features. The results (see Table 4) show that also in this aspect the German tasks stand out: the German tasks have clearly the highest explained variance, in particular for difficulty in reading. Tasks in Swedish and English show similar amount of explained variance for difficulty in solving, while for difficulty in reading, the English tasks show a lower degree of explained variance.

Table 4. Explained variance (R^2) in regression models for prediction of difficulty in reading and difficulty in solving mathematics tasks, with variables measuring all seven linguistic features as independent variables.

	English	**German**	**Swedish**
Difficulty in reading	2.3% (N=62)	22.6% (N=63)	6.7% (N=60)
Difficulty in solving	11.6% (N=83)	20.6% (N=83)	10.7% (N=83)

6. Discussion

6.1 Important differences between different languages

The overarching question in this study is whether the linguistic features of mathematics tasks relate to difficulty in reading and difficulty in solving in different ways for tasks written in English, German, and Swedish. The results show that there in fact are such differences, especially between German and the other two languages, and that these differences primarily exist in relation to difficulty in reading. Specifically, more of the examined linguistic features are significantly correlated to difficulty in reading for the German tasks and the explained variance is also larger for German tasks (for both difficulty in reading and difficulty in solving). Considering that all three languages examined in this study are Germanic languages with many similarities, it is an important finding that differences exist between these related languages. International comparative studies like PISA and TIMSS are based on the assumption that it is possible to compare students' proficiency in mathematics between countries and languages. The results from the present study indicate that such comparisons might be problematic, even for languages that are closely related. The various languages used for comparing students between and within countries are often much more different than the three Germanic languages examined here, which might cause even larger problems with multilanguage assessments.

However, it is also important to note that not all the linguistic features that we examined seem to be related to the students' results. The differences are limited to certain linguistic features, in particular to word length, and neither sentence length nor task length is related to any type of difficulty for any of the languages. It is possible that most of the students in the present study, who all are around 15 years old, are used to these particular linguistic features and therefore, as a group, they are not affected much by them. However, there could be subgroups of students, such as second language learners, that are affected by these features. The effects of different linguistic features in mathematics tasks for such subgroups of students could be more closely examined using, for example, DIF analyses (cf. Heppt, Haag, Böhme, & Stanat, 2015). However, this is beyond the scope of the present study.

6.2 Methodological advantages of the study

The present study contributes with knowledge regarding differences between mathematics tests in different languages, and it has two main advantages in comparison to previous research in the same area. Firstly, it explicitly points to some of the linguistic features that are related to the detected differences. This is an advantage compared to, for example, such studies that we described in the background that use differential item functioning (DIF) analyses to identify problematic tasks, but do not, and often cannot, determine the reasons (e.g., linguistic features) for the differences. Secondly, the present study can methodologically separate necessary from unnecessary reading demands by the use of PCA, resulting in the variable *demand of reading ability* (DRA). This is an advantage compared to, for example, studies that look at correlations between linguistic features and student performance. Such studies can show that some particular linguistic feature is connected to the results for the tasks (e.g., Roe & Taube, 2006) but cannot say whether this connection is necessary, for example, because the linguistic feature is used to express more advanced mathematics, or unnecessary and should therefore be avoided, for example, since the linguistic feature has no relation to mathematics or mathematical ability.

Another methodological conclusion from the results is connected to the fact that not all empirical studies mention which natural language is examined (e.g., see Helwig et al., 1999). The results from this study show that the choice of natural language is an important aspect and should be taken into consideration not only when designing studies but also when analysing data and drawing conclusions, which we discuss more in the subsections below. If we want to be able to draw more general conclusions regarding the relation between linguistic features of tasks and difficulties in reading and solving those tasks, the results show that it is necessary to take into account the variation over different languages, and to examine such relations for many different natural languages, and not generalize results from just one language or a few languages.

6.3 Differences between the languages in relation to the results

One way to explain the different empirical results for the different languages is to compare properties of the languages. The three languages examined in this study are all Germanic languages, which are closely related and similar in many different ways. However, there could still be inherent properties of the languages that cause the differences in how linguistic features are related to difficulties in reading and solving the tasks. One difference between these three languages, addressed in the background, is word order for representing numbers. As mentioned, the number 32 is pronounced *zweiunddreißig* in German, starting with the unit digit, but in English and Swedish the pronunciation order is reversed. However, since our results show that German is different, when compared to English and Swedish, concerning *word length*, it is unlikely that the *word order* for representing numbers would cause these results.

It is possible that structural differences of the languages are causing the differences in how linguistic features are related to different types of difficulties, since the three languages do have some clear structural differences in relation to the linguistic features analysed. One is that the words tend to be longer in German than in the other languages. In this study, a long word is either a word with *more than six letters* or a word with *more than one syllable*. The results show that the fraction of long words (of either type) in the tasks is *negatively* correlated to difficulty in reading in German. In trying to interpret this result, we focus on two different aspects. One is the negative correlation coefficient, and the other is the fact that this relation is valid only for German tasks.

Regarding the first aspect, a negative correlation coefficient in this case means that tasks with a higher unnecessary reading demand (see Figure 1) in general have a smaller proportion of long words. This finding can be seen as somewhat counter-intuitive, and therefore difficult to grasp and explain, but there are several potential explanations. One partial explanation could be that long words in these PISA tasks (in German) are mostly mathematical words or words generally common in mathematics tasks and therefore they do not add to the difficulty in reading, because of how the variable is defined in this study (the variable does not cover the part of reading ability that is included in mathematical ability). In addition, since difficulty in reading is measured using data from both mathematics and reading tasks, it is also possible that relations between the features of these two groups of PISA tasks create this effect. It is also possible that the measures of word length in these particular tasks are indirect indicators of the presence of some other aspect or feature of the task texts and that it is not word length per se that *causes* the difficulty in reading. Another potential explanation is that the result here described as a negative correlation to the fraction of words longer than six letters in fact is a result about a positive correlation to the fraction of words shorter than seven letters, since these descriptions are equivalent. These shorter words might be common words that are always used in texts, such as prepositions and connectives, which might put higher demands of a more general type of reading ability. Since German tends to create compound words for content words, it could be that German, to a larger extent than English and Swedish, has a more distinct separation between the longer content words and the shorter common words. Such a separation might then result in a significant correlation of the type noted in this study. However, it is common also in Swedish to create compound words, as for example seen in Table 1, where both *drip rate* and *drop factor* are translated into compound words in both Swedish and German. Therefore, it is unclear why this effect would only be visible for German. More generally, it is not clear which of these potential explanations are more valid, and further studies are necessary to understand the phenomenon in more detail.

The second aspect is that the significant correlation between long words and difficulty only exists for the German tasks, while the correlation coefficients for the English and Swedish tasks are far from significant. The measure of long words as *multisyllabic* words was chosen since it was used in previous studies that also measured difficulties in solving (student performance) (e.g., Helwig et al., 1999), but it has not been used in studies using the particular variable *demand of reading ability*. It is therefore not possible to directly compare these results with previous studies. The measure *fraction of words longer than six letters* was used in an earlier study where it correlated significantly (and positively) with demand of reading ability for Swedish PISA mathematics tasks from 2003 and 2006 (Österholm & Bergqvist, 2012b). The current results indicate that this relation is not established for Swedish tasks in general, since the results differ between tasks from different years (the tasks used in this study are from PISA 2012). Also, it is possible that the explanation discussed above, that the long words in the PISA tasks are mostly words generally common in mathematics tasks, is only valid for the German tasks. In German it is common to construct long compound words constructed from several short (common and simple) words and this practice can make it possible for students to figure out the meaning of these long words. It is possible that this way of constructing more transparent words therefore could explain these results. A concrete example exists in one PISA mathematics task (M136) in which the German word *Obstgarten* (10 letters, 3 syllables) is used. The word literally means *fruit-garden* and is a translation of the English word *orchard* (7 letters, 2 syllables) that is used in the English version. A German student that is not familiar with this word can figure out its meaning by dividing it into well-known meaningful parts. An English-speaking student who is not familiar with the word orchard cannot do the same.

The way the languages are built and the way their characteristics differ might result in different linguistic features being connected to difficulties in reading in different ways. For example, based on the differences in word length between the languages (see Section 5.1), another type of explanation of our results is that the limit for what is counted as a long word should be different for different languages. The variable *fraction of long words,* where long words means words with more than 6 letters, is significantly correlated to difficulty in reading for tasks in German, but perhaps *long words* has to mean words with more than, for example, 4 letters to be significantly correlating for tasks in English. Such differences between languages could be examined using exploratory studies testing different definitions for different languages, for example, to regard a word as long if it is a certain number of letters or syllables longer than the mean word length in a particular language.

In summary, we have here addressed several possible explanations to the empirical results, all connecting in different ways to properties of the languages. In particular, explanations include the possibility that longer words tend to be important mathematical words to a larger extent in German, or that longer words in German tend to be more transparent when compound words are created, or that the limit of what is relevant to count as a long word is different in different languages. It is beyond the scope of the present study to suggest which explanation is most valid, and more research is needed for this.

6.4 The quality of mathematics tasks

Students trying to solve the tasks examined in this study have different experiences depending on their language. One reason is that the tasks' linguistic features are more important in German than in English and Swedish, concerning primarily difficulty in reading but also difficulty in solving. Similar results have been shown in previous studies, in particular that tasks can function differently in different languages, but such studies seldom reveal any reasons for such differences. However, our study points to one potential cause, namely specific linguistic features of the tasks, which possibly are connected to inherent properties of the different languages, as discussed above. These results are not conclusive, but imply that more research regarding causes of translation issues is needed.

Our results can question the quality of the German tasks, since linguistic features regarding word length are connected to unnecessary reading demands for mathematics tasks. When finding specific linguistic features of tasks that seem to reduce the quality of the tasks, one might suggest to try to remove such a feature, and avoid it when translating tasks. However, since the linguistic feature in this case is word length, which is connected to an inherent property of the language, it is not easy or even possible to remove this type of difficulty. Instead, one needs to be aware of such features of, and differences between, languages when analysing tasks in different languages. However, it is also possible that it is not the length of the words that affects the students' solving of the tasks, but that word length is an indirect measure of some other phenomena that affect the difficulty in reading for the German tasks. Either way, the connection needs to be more thoroughly examined, especially regarding whether the words are mathematical words or not.

6.5 **Conclusions**

Mathematics tasks play an important role within mathematics education and they need to be of high quality and assess mathematical ability. This study has shown that when tasks are translated into different languages, different abilities might be measured, possibly due to linguistic features of the tasks. In particular, PISA tasks in German, when compared with tasks in English and Swedish, show stronger connections between the examined linguistic features of tasks and difficulty in reading and solving the tasks. This result implies that some languages, like German, might introduce more construct-irrelevant variance than others (e.g., see Haladyna & Downing, 2004). Previous research has indeed shown that for PISA, and in other situations when tasks are translated, results from tasks in different languages are not always comparable. We have also shown this in our study, but in addition, we have shown how specific linguistic features could be the reasons for such incomparability. And to some extent, it seems to be possible to relate these linguistic features to inherent properties of the different languages, in particular, concerning general differences in word length. International comparative studies, like PISA, are interesting and important, but should therefore not be overestimated regarding their ability to fairly compare countries. For example, making political decisions regarding the school system or curriculum in a country solely based on results from PISA would be unwise.

Our results have direct implications for research. We agree with Haladyna and Downing (2004), who state that there is a need of more research on how students' language skills affect test performance. More specifically, based on the results in the present study, we argue that there is a particular need for more analyses of different types of languages, when studying and drawing conclusions about the potential relationships between specific linguistic features of tasks and the difficulties students might have when trying to solve them. Therefore, to support further international comparative studies in mathematics (and other subjects), it is important to conduct studies focusing on specific linguistic differences between tasks in different languages and relate these, to issues of students' reading and solving of the tasks, as has been done in the present study.

7. ACKNOWLEDGEMENTS

This study was partially funded by Riksbankens Jubileumsfond, The Swedish Foundation for Humanities and Social Sciences, Grant No. Pl3-0226:1. We want to thank the members of the Advisory Board: Professor Candia Morgan, University College London, England, Professor Sebastian Rezat, Universität Paderborn, Germany, Professor Sara Rezat, Universität Paderborn, Germany, and Professor Caroline Liberg, Uppsala University, Sweden. We also would like to thank our colleagues Dr. Cris Edmonds-Wathen and Moa Eirell for support with analyses.

8. REFERENCES

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). *NAEP Math Performance and Test Accommodations: Interactions with Student Language Background*. CSE Technical Report 536. Los Angeles: University of California. Retrieved from http://www.cse.ucla.edu/products/Reports/TR536.pdf

Abedi, J., Leon, S., Wolf, M. K., & Farnsworth, T. (2008). Detecting test items differentially impacting the performance of ELL students. In M. K. Wolf, J. L. Herman, J. Kim, J. Abedi, S. Leon, N. Griffin, P. L. Bachman, S.M. Chang, T. Farnsworth, H. Jung, J. Nollner, & H.W. Shin (Eds.), *Providing Validity Evidence to Improve the Assessment of English Language Learners* (pp. 55–80). CRESST Report 738. Los Angeles: University of California. Retrieved from http://files.eric.ed.gov/fulltext/ED502627.pdf

Abedi, J., Lord, C., & Plummer, J. R. (1995). *Language background as a variable in NAEP mathematics performance. NAEP TRP Task 3D: Language background study*. Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the Causes of DIF in Translated Verbal Items. *Journal of Educational Measurement, 36*(3), 185–198.

Barwell, R., Barton, B., & Setati, M. (2007). Multilingual issues in mathematics education: introduction. *Educational Studies in Mathematics, 64*(2), 113–119.

Behrman, E. H., & Street, C. (2005). The Validity of Using a Content-Specific Reading Comprehension Test for College Placement. *Journal of College Reading and Learning, 35*(2), 5–21.

Bergqvist, E., Dyrvold, A., & Österholm, M. (2012). Relating vocabulary in mathematical tasks to aspects of reading and solving. In C. Bergsten, E. Jablonka, & M. Raman (Eds.), *Evaluation and comparison of mathematical achievement: Dimensions and perspectives. Proceedings of Madif 8, the eighth Swedish mathematics education research seminar, Umeå, January 24–25, 2012* (pp. 61–70). Linköping, Sweden: SMDF.

Best, R., Ozuru, Y., & McNamara, D. S. (2004). *Self-explaining science texts: Strategies, knowledge, and reading skill.* Paper presented at the Proceedings of the 6th international conference on learning sciences.

Botta, R., Pingree, S., & Hawkins, R. P. (1993). *Does Shorter Mean Easier to Understand? A Study of Comprehension of USA Today Information Stories*. Paper presented at the Annual Meeting of the Association for Education in Journalism and Mass Communication, Kansas City, MO, August 11–14, 1993. Retrieved from http://files.eric.ed.gov/fulltext/ED361678.pdf

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

Chen, F., & Chalhoub-Deville, M. (2015). Differential and long-term language impact on math. *Language Testing, 33*(4), 577–605.

Cowen, C. C. (1991). Teaching and Testing Mathematics Reading. *American Mathematical Monthly, 98*(1), 50–53.

Dryer, M. S., & Haspelmath, M. (2013). *The World Atlas of Language Structures Online*. Retrieved from http://wals.info

Einarsson, J. (1978). *Talad och skriven svenska: Sociolingvistiska studier [Spoken and written Swedish: Sociolinguistic studies]*. Doctoral Dissertation, Lund University, Lund, Sweden.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2*(3–4), 199–215.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential Item functioning identified by expert reviews. *Educational Measurement: Issues and Practice, 29*(2), 24–35.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*(3), 301–321.

Ercikan, K., & Koh, K. (2005). Examining the Construct Comparability of the English and French Versions of TIMSS. *International Journal of Testing, 5*(1), 23–25.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221.

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

Fuson, K. C., & Kwon, Y. (1992). Korean children's understanding of multidigit addition and subtraction. *Child Development*, 491–506.

Geary, D. C., Bow-Thomas, C. C., Liu, F., & Siegler, R. S. (1996). Development arithmetical competencies in Chinese and American children: Influence of age, language, and schooling. *Child development, 67*(5), 2022–2044.

Gerofsky, S. (1999). Genre analysis as a way of understanding pedagogy in mathematics education. *For the Learning of Mathematics*, 36–46.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement, 38*(2), 164–187.

Glazer, S. M. (1974). Is Sentence Length a Valid Measure of Difficulty in Readability Formulas? *Reading Teacher, 27*(5), 464–468.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing, 20*(2), 225–240.

Haag, N., Heppt, B., Roppelt, A., & Stanat, P. (2014). Linguistic simplification of mathematics items: effects for language minority students in Germany. *European Journal of Psychology of Education, 30*(2), 145–167.

Haag, N., Heppt, B., Stanat, P., Kuhl, P., & Pant, H. A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction, 28*, 24–34.

Halliday, M. A. K. (1975). Some aspects of sociolinguistics. In UNESCO (Ed.), *Interactions between linguistics and mathematical education* (pp. 64–73). Retrieved from http://unesdoc.unesco.org/images/0001/000149/014932eb.pdf

Han, Y., & Ginsburg, H. P. (2001). Chinese and English mathematics language: The relation between linguistic clarity and mathematics performance. *Mathematical Thinking and Learning, 3*(2–3), 201–220.

Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L., Mohler, P. P., ... Smith, T. W. (2010). *Survey methods in multinational, multiregional, and multicultural contexts*. Hoboken, NJ: Wiley.

Helwig, R., Rozek-Tedesco, M. A., Tindal, G., Heath, B., & Almond, P. J. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students. *The Journal of Educational Research, 93*(2), 113–125.

Heppt, B., Haag, N., Böhme, K., & Stanat, P. (2015). The Role of Academic-Language Features for Reading Comprehension of Language-Minority Students and Students From Low-SES Families. *Reading Research Quarterly, 50*(1), 61–82.

Jalbert, A., Neath, I., Bireta, T. J., & Surprenant, A. M. (2011). When Does Length Cause the Word Length Effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(2), 338–353.

Jerman, M. (1974). Problem length as a structural variable in verbal arithmetic problems. *Educational Studies in Mathematics, 5*(1), 109–123.

Konior, J. (1993). Research into the construction of mathematical texts. *Educational Studies in Mathematics, 24*(3), 251–256.

Kulm, G. (1971). *Measuring the Readability of Elementary Algebra Using the Cloze Technique*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, N. Y.

Lenzner, T. (2014). Are Readability Formulas Valid Tools for Assessing Survey Question Difficulty? *Sociological Methods & Research, 43*(4), 677–698.

Lepik, M. (1990). Algebraic word problems: Role of linguistic and structural variables. *Educational Studies in Mathematics, 21*(1), 83–90.

Leung, F. K. (2014). What can and should we learn from international studies of mathematics achievement? *Mathematics Education Research Journal, 26*(3), 579–605.

Lin, P.-Y. (2012). *Setting accommodation and item difficulties.* Doctoral Dissertation, University of Toronto.

Lithner, J., Bergqvist, E., Bergqvist, T., & Boesen, J. (2010). Mathematical competencies: A research framework. In C. Bergsten, E. Jablonka, & T. Wedege (Eds.), *Mathematics and mathematics education: Cultural and social dimensions. Proceedings of MADIF 7, The Seventh Mathemat-*

ics Education Research Seminar, Stockholm, January 26–27, 2010 (pp. 157–167). Linköping, Sweden: SMDF.

Liu, Y., Lin, D., & Zhang, X. (2016). Morphological awareness longitudinally predicts counting ability in Chinese kindergarteners. *Learning and Individual Differences, 47*, 215–221.

Marmurek, H. H. C. (1988). Reading Ability and Attention to Words and Letters in Words. *Journal of Reading Behavior, 20*(2), 119–129.

McKenna, M. C., & Robinson, R. D. (1990). Content literacy: A definition and implications. *Journal of Reading, 34*(3), 184–186.

Morgan, C., Craig, T., Schuette, M., & Wagner, D. (2014). Language and communication in mathematics education: An overview of research in the field. *ZDM - the International Journal on Mathematics Education, 46*, 843–853.

NCTM. (2000). *Principles and standards for school mathematics* (Vol. 1). National Council of Teachers of Mathematics.

Niss, M., & Højgaard, T. (Eds.). (2011). *Competencies and mathematical learning: Ideas and inspiration for the development of mathematics teaching and learning in Denmark*. Roskilde, Denmark: Roskilde University.

Norgaard, H. L. (2005). *Assessing linguistic, mathematical, and visual factors related to student performance on the Texas assessment of knowledge and skills, eighth grade mathematics test.* Doctoral Dissertation, University of North Texas.

Nuerk, H.-C., Weger, U., & Willmes, K. (2005). Language effects in magnitude comparison: Small, but not irrelevant. *Brain and Language, 92*(3), 262–277.

OECD. (2009). *PISA 2009 assessment framework: Key competencies in reading, mathematics and science.* Paris, France: OECD.

OECD. (2010). *Translation and Adaption Guidelines for PISA 2012*. Budapest, Hungary: OECD.

OECD. (2014). *PISA 2012 Technical Report*. Paris, France: OECD Publishing.

Perfetti, C. A. (1969). Lexical density and phrase structure depth as variables in sentence retention. *Journal of Verbal Learning and Verbal Behavior, 8*(6), 719–724.

Pirkola, A. (2001). Morphological typology of languages for IR. *Journal of Documentation, 57*(3), 330–348.

Pixner, S., Moeller, K., Hermanova, V., Nuerk, H. C., & Kaufmann, L. (2011). Whorf reloaded: Language effects on nonverbal number processing in first grade - A trilingual study. *Journal of Experimental Child Psychology, 108*(2), 371–382.

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.

Roth, W.-M., Ercikan, K., Simon, M., & Fola, R. (2015). The assessment of mathematical literacy of linguistic minority students: Results of a multi-method investigation. *Journal of Mathematical Behavior, 40*, 88–105.

Roth, W.-M., Oliveri, M. E., Sandilands, D. D., Lyons-Thomas, J., & Ercikan, K. (2013). Investigating Linguistic Sources of Differential Item Functioning Using Expert Think-Aloud Protocols in Science Achievement Tests. *International Journal of Science Education, 35*(4), 546–576.

Rothkopf, E. Z., & Kaplan, R. (1972). Exploration of the effect of density and specificity of instructional objectives on learning from text. *Journal of Educational Psychology, 63*(4), 295–302.

Schleppegrell, M. J. (2007). The Linguistic Challenges of Mathematics Teaching and Learning: A Research Review. *Reading & Writing Quarterly, 23*(2), 139–159.

Sfard, A. (1991). On the dual nature of mathematical conceptions: Reflections on processes and objects as different sides of the same coin. *Educational Studies in Mathematics, 22*(1), 1–36.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(2), 105–126.

Shanahan, T., & Shanahan, C. (2008). Teaching Disciplinary Literacy to Adolescents: Rethinking Content-Area Literacy. *Harvard Educational Review, 78*(1), 40–59.

Sigurd, B., Eeg-Olofsson, M., & Van Weijer, J. (2004). Word length, sentence length and frequency – Zipf revisited. *Studia Linguistica, 58*(1), 37–52.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5 ed.). Boston, MA: Allyn and Bacon.

White, S. (2012). Mining the text: 34 text features that can ease or obstruct text comprehension and use. *Literacy Research and Instruction, 51*(2), 143–164.

Wichmann, S., Holman, E. W., & Brown, C. H. (2016). *The ASJP Database (version 17)*. Retrieved from http://asjp.clld.org

Wolf, M. K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment, 14*(3–4), 139–159.

Ziegler, J. C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., ... Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. *Psychological Science*, *21*(4), 551–559.

Ziegler, J.C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, 131, 3–29.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF).* Ottawa, Canada: National Defense Headquarters.

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

Österholm, M., & Bergqvist, E. (2012b). 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, E. Jablonka, U. T. Jankvist, A. Ryve, P. Valero, & K. Wæge (Eds.), *Proceedings of Norma 11, The Sixth Nordic Conference on Mathematics Education in Reykjavík, May 11–14, 2011* (pp. 661–670). Reykjavík, Iceland: University of Iceland Press.