# Equivalence in multilanguage mathematics assessment

FRITHJOF THEENS, EWA BERGQVIST AND MAGNUS ÖSTERHOLM

When mathematics tasks are used in multilanguage assessments, it is necessary that the task versions in the different languages are equivalent. The purpose of this study is to deepen the knowledge on different aspects of equivalence for mathematics tasks in multilanguage assessment. We analyze mathematics tasks from PISA 2012 given to students in English, German and Swedish. To measure formal equivalence, we examine three linguistic features of the task texts and compare between language versions. To measure functional equivalence, a *Differential item functioning* (DIF) analysis is conducted. In addition, we examine statistically if there is a relation between DIF and the differences regarding linguistic features. The results show that there is both DIF and differences regarding the linguistic features between different language versions for several PISA tasks. However, we found no statistical relation between the two phenomena.

Multilanguage assessment in mathematics, that is, giving a mathematics test to students in different languages, is common and serves different purposes. International comparative assessments like the *Programme for international student assessment* (PISA) are used to compare and evaluate educational systems in different countries, by testing the skills and knowledge of students. There are also national multilanguage assessments that, for example, serve as a basis to examine the curriculum of different provinces in multilingual countries (e.g. Pan-Canadian assessment program, O'Grady, 2018) or that focus on assessing students, for example, when used for students in English classes in Sweden (e.g. the national standardized test in Sweden, Swedish National Agency for Education, 2018).

**Frithjof Theens**, *Umeå University*
**Ewa Bergqvist**, *Umeå University*
**Magnus Österholm**, *Umeå University and Mid Sweden University*

Theens, F., Bergqvist, E. & Österholm, M. (2023). Equivalence in multilanguage mathematics assessment. *Nordic Studies in Mathematics Education*, 28 (1-2), 7–29.

When a multilanguage assessment is administered to different language groups, the issue of equivalence between the language versions has to be taken into consideration. A lack of equivalence between language versions might jeopardize the validity of the assessment if the inequivalence favors one group of students. Since the results from international comparative assessment, such as PISA, are used at a political level, for policy change and implementation, a lack of validity in such assessments can be detrimental to educational systems in different countries. When multilanguage assessments are used within one country, a lack of validity caused by an inequivalence between language versions could cause unequal opportunities for different student groups if such assessments are used for admissions to certain educational levels.

However, the concept of equivalence can encompass many different issues. For example, Arffman (2010) describes that two language versions of a text can be equivalent (or inequivalent) regarding text properties, such as contents, semantics, style, register, or formal-aesthetic aspects. Arffman also addresses an aspect of pragmatic equivalence, "which concentrates on the reader and refers to equivalence of effect" (p. 40). At an overarching level, it is therefore possible to distinguish between two types of perspectives on test equivalence (see also Pym, 2010): a *formal perspective* that focuses on properties of the texts (by comparing a translated version with a source version or with other translated versions) and a *functional perspective* that focuses on the effects the texts have on readers, e.g. concerning aspects of comprehension.

What makes the concept of equivalence complex is the fact that these different perspectives on equivalence can sometimes be contradictory, since it is not possible to have perfect equivalence in all aspects when translating (Koller, 2011). Arffman (2010, p. 46) gives one example, when translators are "following too closely the original texts" (focusing on high degree of formal equivalence) and thus making the translated text "partly unintelligible" (reducing the functional equivalence). Therefore, decisions have to be made concerning what to prioritize. When focusing on student assessment, the functional perspective can be seen as the more important perspective, since in the context of assessment it is essential to achieve a "similar level of difficulty or comprehensibility between the source and target texts" (Arffman, 2010, p. 40). The functional perspective on equivalence is closely related to the two core concepts of reliability and validity (cf. Arffman, 2010). At the same time, the formal perspective on equivalence cannot be ignored, which, for example, is highlighted in the guidelines for translation of PISA tasks (OECD, 2010, p. 8): "The translation must not be literal to the point that it sounds awkward, but neither should it deviate too far from the source version,

which would likely affect the functioning of the assessment items in unexpected ways."

Due to these complexities of different perspectives on equivalence, it is important to know if and how the different perspectives are related to each other. Is it possible to find a specific lack of equivalence from one perspective as a reason for a lack of equivalence from another perspective? For example, can we find certain linguistic differences between language versions of a task that are related to differences in difficulty or comprehensibility for students when solving the task? Answers to these types of questions are important since they can enable the use of multilanguage assessment in a valid and reliable manner. Therefore, the purpose of this study is to *deepen the knowledge on relationships between different aspects of equivalence for mathematics tasks in multilanguage assessment*. We quantitatively analyze mathematics tasks in English, German and Swedish regarding whether differences between the tasks concerning certain linguistic properties (formal perspective) are related to differences in the level of difficulty for students to solve these tasks (functional perspective). Our analyses complement previous research that has been more qualitative and exploratory. The chosen three languages are all Germanic languages and have many features in common but have nevertheless shown differences regarding properties that affect reading and solving mathematics tasks (Bergqvist et al., 2018).

## Equivalence of tasks in multilanguage assessment

The tasks used in multilanguage assessments have to be translated from one or several source versions (for PISA-tasks from English and French) into one or several different languages. The translated versions should all be as equivalent as possible to give comparable and useful results of the assessment. A main requirement is, of course, that the translation is free from obvious errors, but it is impossible to create different language versions of a task including text that are equivalent from all relevant perspectives. Below we present and discuss two main perspectives on equivalence – functional and formal perspectives (cf. Pym, 2010).

### *A functional perspective on equivalence*

The functional perspective on equivalence focuses on how comparable the effects on the students are, usually by comparing test results, when focusing on assessments. From this perspective, different language versions of a task have to "measure the same concept at a comparable level of difficulty" (Arffman, 2010, p. 39). If this is achieved, the test results

can be considered both reliable and valid across the different language versions and equivalence (from this perspective) is reached. If this type of equivalence is not reached, a task is said to *function differently* in the different language versions.

Differential item functioning (DIF) is a common method for examining the existence and degree of inequivalence from a functional perspective (Zumbo, 1999). A task shows DIF if test takers from different groups have different probabilities to answer the task correctly despite having the same underlying ability that the test intends to measure (AERA, 2014). This difference in probability "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). The occurrence of DIF can have different reasons and groups used in DIF analysis can be based on, for example, gender, socio-economic status, or language proficiency. However, DIF can also be used to compare groups that have taken different language versions of a test. Empirical studies show that many translated tasks display DIF between different language versions (Hopfenbeck et al., 2017), in some cases up to 79 % of tasks in a test (Ercikan & Koh, 2005).

### *A formal perspective on equivalence*

The formal perspective on equivalence focuses on comparisons of text properties of different language versions of a task. To be equivalent from this perspective, language versions have to be related in a way that they share amongst others semantic, formal-aesthetic and textual features (Arffman, 2010; Solano-Flores et al., 2009). Translated texts in general have particular features that separate them from non-translated texts, for example, passive voice is used more often in English translated texts than in non-translated ones (Volansky et al., 2013; Xiao & Yue, 2009). A translated text in a particular language can be seen to be written in "a dialect of that language, known as 'translationese' " (Volansky et al., 2013, p. 98), which produces some degree of inequivalence between language versions. When analyzing language versions of assessment tasks, the formal perspective is usually not focused on separately. Instead, it is often addressed in relation to the functional perspective, since the latter is seen as more important for such tasks (see Arffman, 2010).

### *Potential relations between formal and functional equivalence*

If two language versions of a task are different regarding linguistic features, they are less equivalent from the formal perspective. If one language version becomes more difficult to solve because of a special linguistic

feature, the functional equivalence also decreases, and the different task versions could display DIF. In earlier studies, certain linguistic differences have been identified as reasons for DIF between language versions of translated tasks, for example:

- Word difficulty, meanings/connotations, sentence complexity, grammatical form and idiomatic relationships (Allalouf, 2003; Allalouf et al., 1999).
- Unfamiliar terms or expressions, sentence complexity, negations and logical relations (Roth et al., 2013).
- Word difficulty, grammatical structure and contextual meaning (Huang et al., 2014).
- Key vocabulary, sentence complexity and clarity (Ercikan et al., 2004).

The standard method to find relations between linguistic features and DIF used in these types of studies is to first calculate DIF for a set of tasks and then have linguistic experts perform exploratory analyses of the tasks to find possible reasons for DIF. However, the exploratory nature of this method also has its limitations. For example, experts are not always able to identify task features that distinguish tasks that exhibit DIF from tasks that do not exhibit DIF. Experts sometimes also think that DIF is reversed (i.e., in favor of the other language) compared to results from the DIF analysis (Roth et al., 2013). Similarly, Ercikan et al. (2010) found that in many cases (9 of 20 tasks), analyses of students' thinking processes when working with the tasks could not confirm the results from experts' analyses.

Thus, previous empirical research on reasons for DIF has been exploratory and qualitative, through a reliance on expert judgments, which do not always appear to be that reliable. Therefore, in the present study, we investigate quantitatively if and how some specific linguistic differences are related to occurrence of DIF in translated tasks. The benefit of this approach is a possibility to analyze many items in several languages focusing on specific properties, which could reveal if there are any more general patterns concerning such properties. Our approach could therefore complement existing research that has been more exploratory and qualitative.

To make a more in-depth quantitative analysis of linguistic features, we delimit our study to a few specific features. For this analysis, we also had to operationalize each feature in a quantitative way, which excludes some features located in previous, more exploratory, research.

The linguistic features chosen in the present study are *voice*, *grammatical person* and *sentence structure*. Below, we give a brief summary of previous research regarding each of these features concerning potential relations to functional equivalence and argue for the relevance to examine these features in our study.

*Voice* is the grammatical expression of the relationship between the action, subject and object in a text (Chicago manual of style, 2003, p. 176). A sentence can be written in either active or passive voice. In active voice, the subject of the sentence is the agent who is acting ("The dog eats the cake"). In passive voice, the subject of the sentence is the target who is being acted on ("The cake is eaten by the dog"). Previous research has sometimes shown connections between voice and task difficulty or readability, although the relation is not completely clear, which makes it relevant to examine in more research studies. Passive sentences seem to take longer time to read (Bostian, 1983; Forster & Olbrei, 1973) and very young children (19–38 months) are better at acting out active than passive sentences (Villiers & Villiers, 1973). Still, passive voice does not always affect readability negatively, since "appropriately used passives may actually improve readability" (Allan, 2009, p. 190). However, Abedi, Lord and Plummer (1997) found that mathematics tasks with active voice were significantly easier for students in average mathematics classes, than the same tasks written in passive voice. A difference in the use of active and passive voice between the language versions of a task could therefore give an advantage for one of the language groups. Therefore, we include voice as a feature to examine.

*Grammatical person* refers to the distinction between the first person (e.g. I and we), the second person (e.g. you) and the third person (e.g. she and they). A text using second person addresses the reader directly and is by that more personal than a text using third person. Empirical research shows a connection between personalization of mathematics tasks and higher success rates when students are solving the tasks (Davis-Dorsey et al., 1991; Ross & Anand, 1987). In these studies, students in grades 2–6 benefited from personalization of tasks. Making the task text more personal by changing from third to second person might enhance student performance on the task also for older students. On the other hand, the pronoun "you" can in English also be used in an impersonal, generic way (e.g. "In the USA, you always eat turkey on Thanksgiving"). In German and Swedish, there is a special impersonal pronoun "man" that is used for generalization. Therefore, two different ways to translate an English text using "you" to German or Swedish are possible. Either "you" is interpreted in a more personal way and then translated with "du" (second person singular) or it is interpreted in an impersonal way and then translated

with "man" (third person singular). Thus, based on previous research showing a potential effect on student performance and on the differences between languages, we include grammatical person as a feature to examine in this study.

*Sentence structure* refers to how a sentence is built of one or several clauses. The simplest structure is a sentence containing just one main clause, for example, "The dog eats the cake". But sentences are often more complex. They can be built of one main clause in combination with one or several subordinate clauses, for example, relative or conditional clauses (e.g. "The dog eats the cake, which he stole from the cat"). Sentences can also be compounds of two or more main clauses (e.g. "The dog stole the cake, he ate it, and the cat got angry"). The presence of subordinate clauses in mathematics tasks showed a correlation to lower scores on the tasks and longer time needed to solve the task (Lord, 2002). If the sentence structures differ between language versions of a task, the language version with simpler structure could give an advantage for the corresponding language group. Therefore, we include sentence structure as a feature to examine in this study.

## Purpose and research questions

The purpose of this study is to *deepen the knowledge on relationships between different aspects of equivalence for mathematics tasks in multilanguage assessment*. This purpose is reached by examining how linguistic features differ between language versions of tasks and for which tasks there is DIF between language versions. In addition, we examine the relation between these variables statistically. The study examines PISA mathematics tasks in English, German and Swedish. We address the following research questions:

1 To what extent do the tasks display DIF between different language versions?

2 How much do the linguistic features differ between the different language versions?

3 How is DIF related to differences concerning the linguistic features?

## Method

To answer the research questions, the analysis is carried out in three steps for each pair of language versions of the tasks, that is, three steps for English compared to German tasks, the same three steps for English to Swedish and again the same three steps for German to Swedish.

During the first step, we determine whether the language versions of each task functioned statistically different by calculating the level of *Differential item functioning* (DIF) between them. The results from this step are used to answer the first research question. In the second step, we measure how much the language versions differ regarding each linguistic feature. This analysis answers the second research question. The last step answers the third research question by analyzing relations between the *level of DIF* and *the amount of linguistic differences* through regressions. In the following sections, the data selection and the three steps of the analysis are explained in more detail.

### *Data selection*

In this study, we include all mathematics tasks from the PISA 2012 assessment used in the USA (English), in Germany (German) and in Sweden (Swedish). The focus of this study is not on PISA but there are two main reasons why we chose to use PISA tasks. Firstly, the translations of the tasks were made by professionals that follow a strict procedure, which minimizes the risk of pure translation errors (OECD, 2010). Secondly, PISA tests are taken by a large number of students in different languages and all students work with many tasks. The large amount of data makes it possible to calculate Differential item functioning (DIF) and also to quantify the linguistic features of the tasks in a reliable way. All student results on the tasks we needed for the analysis are available on the PISA website (OECD, 2012).

In PISA 2012, there are a total of 84 mathematics tasks. Some of these tasks are related to each other through a common introductory text. Each student participating in PISA was assigned one of 13 booklets containing between 11 and 37 of the mathematics tasks. Each task was included in four different booklets. That is, each student only worked with a selection of the 84 tasks, and we therefore calculate DIF between different language versions of tasks for one booklet at a time (see details in the next section). For each booklet, there are results from 300–400 students in each language group. For one task in one booklet there were very few answers from the Swedish students. Therefore, this task was excluded from the analysis in this booklet when doing comparisons with the Swedish students.

### *Calculating level of DIF between language versions*

For the calculation of Differential item functioning (DIF) between the different language versions of the tasks, we analyzed the language versions pairwise. We used the Mantel-Haenszel method, which is

commonly used for DIF-detection (Ferne & Rupp, 2007). This method can be used to analyze the results on a task from two different student groups by comparing the amount of correct answers[1] of students from the two groups with equal mathematical ability, which is determined by the total score on the assessment (Allen & Donoghue, 1996). The result of the analysis is the level of DIF on a scale (A = negligible, B = intermediate, C = large)[2] and a positive or negative sign that indicates which group is favored. For example, positive DIF on level B between the German and Swedish version of a task indicates that the task is slightly easier for German students than for Swedish students, and vice versa for negative DIF.

Since each task was included in four different booklets containing different tasks, we performed the DIF analysis for each task for each booklet separately. In some cases, the statistical assumption regarding homogeneity of the odds ratio, necessary to perform a Mantel-Haenszel-analysis (see e.g. Allen & Donoghue, 1996), was violated. When this happened, DIF was calculated for the task only with the booklets where this assumption was not violated. Using fewer than four booklets is possible and unproblematic since the information in the data for each booklet is sufficient to calculate DIF (see e.g. Çikrikçi Demirtaşli & Ulutaş, 2015). For each task, the analysis thus generated at most 4 values of DIF between each pair of language versions. For each pair of language versions, we took the mode of these values as the DIF value for the task, as exemplified in table 1. If the result was bi- or multimodal, the DIF value was set to the lower level of DIF to avoid drawing too strong conclusions (e.g. see Dorans & Holland, 1992).

For most of the 84 tasks, all four booklets containing the task could be included in the analysis, but for between 19 and 22 tasks in each of the three different language comparisons, not all four booklets could be used due to violation of the statistical assumption. When DIF was found in booklets of a task, the direction of DIF was always in favor of the same language in all booklets, which is seen as a sign of reliability in the method.

Table 1. *Example of calculations of levels of DIF for a task*

| Task X | Booklet 1 | Booklet 2 | Booklet 3 | Booklet 4 | Level of DIF (mode) |
|---|---|---|---|---|---|
| English-German | B + | A | C + | C + | C + |
| English-Swedish | no result | A | B + | no result | A |
| German-Swedish | B - | no result | A | B - | B - |

Since we analyze three languages, it is possible to do a multi-group DIF analysis, with the purpose to address the issue of parallel analyses on several languages. The main problem with such parallel analyses is the repeated tests of statistical significance, causing potential problems with Type I errors (cf. Penfield, 2001), which can create unwarranted conclusions from data. However, the structure of our data set is not suitable for a more direct consideration of the multi-group DIF analysis. In particular, we do the DIF analysis in 3–4 different booklets for each item and based on these we create a measure of the level of DIF for this item. Therefore, instead of doing a more direct adjustment in the DIF analysis (e.g. by using Bonferroni corrections, cf. Penfield, 2001), we do other things in order not to draw unwarranted conclusions: We use the mode of the DIF level from the different booklets and we use the lower DIF level in the case when there are two modes.

### *Calculating amount of linguistic differences*

The second step of the analysis is to calculate the pairwise differences between the three language versions of each task regarding each linguistic feature.

First, we marked every *sentence* where there was a difference between the language versions regarding any of the linguistic features. In particular, we marked if one version used active voice and the other passive voice, if one used second person (”you”) and the other third person (”he”, ”she”, ”it”, ”they”)[3], and if one version used a simple sentence consisting of just one main clause and the other a complex or compound sentence consisting of several main and/or subordinate clauses.

Second, the language versions were compared pairwise and the amounts of differences for each feature were calculated for each task. For all three linguistic features, we used positive and negative values for the differences between language versions to make visible which language version had more of the feature. For example, if there were two sentences in the English version of a task that used passive voice where the Swedish version used active voice, the sum for this task was +2. If it was the opposite, the sum was -2. If there was no difference in the use of passive voice the value was zero. Also, if a task would include one sentence with +1 and another with -1, the task would have the sum 0, even though there is a difference. However, this never occurred in our data. Differences regarding grammatical person and sentence structure were calculated in a similar way. The fact that we count the *number* of occurrences is based on the following reasoning. If a certain feature of a sentence can make this sentence more difficult to understand, then multiple occurrences can

make a larger part of a text more difficult to understand. Therefore, larger differences have the potential to create clearer or larger DIF.

A general note is that some PISA tasks share a common introductory text. The introductory texts differ in length from just some words up to several sentences. Any existing introductory text is included in the analysis of each task since this text might be needed by the students to understand and solve the task.

### *Relations between differences in linguistic features and DIF*

For each pair of language versions, we performed a regression analysis with *the amount of differences in each of the three linguistic features* as independent variables and *the level of DIF* between the language versions as dependent variable. To make the results easy to interpret, we defined the sign of the variables so that any *positive* relation would mean that what we suspected to be more difficult text in one language was connected to favoring students taking the test in the other language. More specifically (see table 2), positive values for linguistic differences are related to the first language in the pair of languages, concerning more of the variant of the linguistic feature that was expected to cause difficulties (i.e., passive voice, third person or complex sentence structure). Within a language pair, a positive sign for DIF then means a favoring of the students taking the test in the second language in the pair of languages.

For ordinal data with few categories, as we have with five categories when measuring DIF, there is not much to gain in using ordinal regression when compared to ordinary least squares regression, if the variable

Table 2. *Sign of the variables for the analyses of relations between differences in linguistic features and the value of DIF*

| Versions | Sign | Voice | Grammatical person | Sentence structure | DIF |
|---|---|---|---|---|---|
| ENG-GER | + | ENG more passive | ENG more 3rd person | ENG more complex | favor GER |
| | - | GER more passive | GER more 3rd person | GER more complex | favor ENG |
| ENG-SWE | + | ENG more passive | ENG more 3rd person | ENG more complex | favor SWE |
| | - | SWE more passive | SWE more 3rd person | SWE more complex | favor ENG |
| GER-SWE | + | GER more passive | GER more 3rd person | GER more complex | favor SWE |
| | - | SWE more passive | SWE more 3rd person | SWE more complex | favor GER |

is not skewed (Taylor et al., 2006). Skewness in our data is essentially non-existent, with values of 0.0; -0.16; and 0.06 for the three pairs of languages, and there is also no problem with multicollinearity, with all VIF values smaller than 1.02. Therefore, we use ordinary least squares regression in our analyses.

For a variable with five categories, a 14 % larger sample size is needed when compared to a continuous variable for the same level of statistical power (Taylor et al., 2006). To detect medium effects with a power of 80 %, when using three independent variables and $p<0.05$, we would need a sample of 87. Our sample is 84 mathematics tasks, making it possible to detect medium to large effects.

## Results

In this section, the results of the analyses are presented, and the three research questions are answered in three separate sections.

### *Tasks with DIF between language versions (RQ1)*

The first research question concerns to which extent the 84 mathematics tasks of the PISA 2012 assessment display DIF between the English, German, and Swedish versions. 43 of the tasks did not display DIF for any pair of languages examined. For the 41 tasks displaying DIF, table 3 shows the number of tasks displaying DIF for each combination of language versions and in favor of each language at both intermediate and high level of DIF. For example, in the upper-left corner of table 3, we see that 3 tasks display large DIF between the English and the German version, in favor of the English version.

There were differences in the pairwise comparison of the English, German, and Swedish versions, both regarding the number of tasks displaying DIF and the distribution of favor to the different language versions. Fewest tasks displaying DIF occur when comparing the German

Table 3. *Number of tasks among 84 mathematics tasks displaying DIF between different language versions*

| | ENG-GER | | ENG-SWE | | GER-SWE | |
|---|---|---|---|---|---|---|
| Version favored by DIF | ENG | GER | ENG | SWE | GER | SWE |
| Large DIF | 3 | 3 | 5 | 5 | 1 | 1 |
| Intermediate DIF | 9 | 9 | 5 | 12 | 5 | 7 |
| Total | 12 | 12 | 10 | 17 | 6 | 8 |
| | 24 | | 27 | | 14 | |

and the Swedish version. When comparing German and Swedish, 17 % (14 of 84) of the tasks display DIF and DIF is quite evenly distributed favoring each of the language versions. When comparing the English and the German version, DIF occurs in 29 % (24 of 84) of the tasks, and the favor is evenly distributed. A different pattern appears when comparing the English and the Swedish versions of the tasks. Here, about 32 % (27 of 84) of the tasks display DIF and more tasks than in the other comparisons display large DIF – five in favor of each language version. Furthermore, the 17 tasks displaying moderate DIF show a clear majority in favor of the Swedish version. A complete list of all tasks displaying DIF between the different language versions can be found in the appendix.

### *Differences regarding linguistic features between language versions of tasks (RQ2)*

The second research question concerns how much the different language versions of the PISA tasks differ regarding voice, grammatical person, and sentence structure. Table 4 shows the differences in these linguistic features between English, German, and Swedish. In some tasks, some types of differences occur several times. For example, there are tasks where active voice is used several times in one language version and where passive voice is used in the other language version. Among the 84 tasks, some tasks do not have any differences at all between the language versions: 28 tasks when comparing English and German, 27 tasks when comparing English and Swedish, and 25 tasks when comparing German and Swedish. On the other hand, some tasks differ in several features. As

Table 4. *Number of occurrences of differences in linguistic features for 84 mathematics tasks*

| | ENG-GER | | ENG-SWE | | GER-SWE | |
|---|---|---|---|---|---|---|
| | ENG | GER | ENG | SWE | GER | SWE |
| Passive voice in this version, active in the other | 18 | 17 | 15 | 31 | 13 | 30 |
| Third person in this version, second person in the other | 6 | 11 | 2 | 8 | 5 | 6 |
| Complex or compound sentence in this version, simple in the other | 19 | 13 | 6 | 18 | 8 | 26 |
| Total number of occurrences of differences | 43 | 41 | 23 | 57 | 26 | 62 |
| | 84 | | 80 | | 88 | |

most, one task has 10 differences in voice and one in sentence structure when comparing the English and the Swedish version.

In general, the least amount of differences concern the use of grammatical person, while most differences occur in the use of voice. For example, between the English and the Swedish versions there are altogether 46 occurrences of difference in the use of voice (15 occurrences with passive voice in English and active in Swedish and 31 with passive voice in Swedish and active in English), and altogether 10 occurrences of difference in the use of grammatical person. Furthermore, the types of linguistic features that are potentially associated with higher complexity (passive voice, third person, and complex/compound sentence) are overall more frequently used in the Swedish version, when compared both with the English and with the German versions.

### *Relations between differences in linguistic features and DIF (RQ3)*

The third research question concerns if the differences in the linguistic features between the language versions relate to DIF between the versions. As shown in table 5, there are no statistically significant relations between *amount of differences* in the linguistic features and *level of DIF* in these tasks.

Table 5. *Regression models with level of DIF as dependent variable and the amount of differences in voice, grammatical person, and sentence structure as independent variables, for 84 mathematics tasks*

| | ENG-GER | ENG-SWE | GER-SWE |
|---|---|---|---|
| Model fit | $R^2 = 0.024$ ($p = 0.575$) | $R^2 = 0.025$ ($p = 0.570$) | $R^2 = 0.007$ ($p = 0.906$) |
| Voice | $\beta = -0.119$ ($p = 0.284$) | $\beta = -0.137$ ($p = 0.219$) | $\beta = -0.082$ ($p = 0.469$) |
| Grammatical person | $\beta = 0.068$ ($p = 0.542$) | $\beta = 0.063$ ($p = 0.575$) | $\beta = 0.002$ ($p = 0.986$) |
| Sentence structure | $\beta = -0.088$ ($p = 0.430$) | $\beta = -0.042$ ($p = 0.708$) | $\beta = 0.027$ ($p = 0.808$) |

## Discussion and conclusions

This study focuses on different aspects of equivalence of language versions of mathematics tasks. The results showed that there indeed occurs Differential Item Functioning (DIF) between different language versions of PISA tasks in English, German and Swedish (see table 3), including

several instances of large DIF. In addition, there are differences regarding voice, grammatical person, and sentence structure between the different language versions of the tasks (see table 4). We found no statistical relation between the two phenomena, that is, we could not detect any relation between DIF and differences regarding these particular linguistic features. Here we discuss different possible explanations of our results as well as implications for future research and policy.

### *Differences between the tasks*

It might be a bit surprising that we found quite many linguistic differences between the language versions, since PISA has extensive routines for the translation of tasks (OECD, 2010). However, all languages, even those closely related, have different inherent properties, concerning vocabulary (Wichmann et al., 2016) and structural properties (Dryer & Haspelmath, 2013), which makes it impossible to perform perfect translations. Research also shows that translated texts in general have particular features that separate them from non-translated texts; that they are written in "translationese" (Volansky et al., 2013). Therefore, it is reasonable that our results reflect ambitions among translators to avoid "translationese" and instead create texts that are more "normal" within a language, which can create such differences between language versions as seen in this study.

Furthermore, our results show that many tasks display DIF, between 17 and 32 % depending on which languages that are compared. This is in line with previous research, for example, another study of PISA tasks showed that 18–46 % of the tasks displayed DIF (Grisay & Monseur, 2007). Potential reasons for DIF are discussed below, concerning relations between functional equivalence (DIF) and formal equivalence (linguistic features).

### *Absence of relation between DIF and the linguistic differences*

One possible reason that the DIF we found was not related to the particular linguistic features that we examined, is that it simply does not matter for students of this age (about 15 years old) whether a task is written in passive or active form, is more or less personal, or contains more or less complex sentences, at least not to the extent it happens in PISA mathematics tasks. Even if these tasks include verbal text to a larger extent than what is common in many mathematics textbooks, the texts are still not very long. There are also some indications in previous research that the possible problems that children might have with these linguistic

features decrease with school year. This could be the case on a general level, that is, that the relation between reading comprehension and mathematics is weaker at higher levels. For example, when Hickendorff (2013) investigated the role of reading comprehension for students in grade 1 and 3 solving mathematics tasks, she found that reading comprehension had lower impact in higher grades. However, other studies show correlations between reading comprehension and mathematics performance of similar magnitudes in different age groups (e.g. Aiken, 1972).

Still, there could be an age effect more specifically for the linguistic features that we have examined in this study. For example, earlier research showed that personalization of mathematics tasks could enhance students' performance on the tasks but focused on younger students than in the present study: Davis-Dorsey et al. (1991) looked at grade 2 and 5, and Ross and Anand (1987) studied grade 5 and 6. It is possible that a personalization does not affect eighth-grade students in the same way. Similarly, the use of passive voice seems not to be problematic for students of this age, as indicated in a study where students commented on their work with mathematics PISA tasks (Theens, 2019). However, more research is needed that examines several age groups of students, to determine if age indeed is a relevant aspect concerning the effects of different linguistic features of mathematics tasks.

It is also possible that there is a relation between DIF and one or more of the linguistic features, but that we cannot detect it using the operationalizations we chose here. There are in fact differences in the operationalizations used in our study compared with other studies showing connections between linguistic features and aspects of reading or solving texts or tasks. For example, in the studies of Davis-Dorsey et al. (1991), and Ross and Anand (1987), the names in the tasks were exchanged for names of the students' friends, to make the tasks more personal. This operationalization could capture a different type of personalization than the changes from third to second person that we study here. In our study, we used pre-existing tasks from PISA, where we could not control the variation of linguistic features. For example, the specific type of variation examined by Davis-Dorsey et al. (1991), and Ross and Anand (1987) was not possible to examine with our data.

However, it is also possible that there is a relation between differences regarding linguistic features and functional equivalence, but that this relation is more complex than what can be captured with the statistical analyses used in our study. For example, a high degree of equivalence from one perspective, concerning linguistic features, could be associated with a *low* degree of equivalence from another perspective, concerning DIF. This relation is discussed by some researchers, for example, by noting that

it is not possible to have perfect equivalence in all aspects when translating (Koller, 2011). We have not found empirical research that focuses on this type of relation, but Arffman (2010, p. 46) gives one example, when translators are "following too closely the original texts" (focusing on high degree of formal equivalence) and thus making the translated text "partly unintelligible" (reducing the functional equivalence). That is, a linguistic feature like, for example, passive voice might be more common, and maybe thereby less difficult, in some languages than in others. Other types of analyses would be needed to detect this more complex type of relation between the different perspectives on equivalence.

In the present study, we assumed that particular aspects of the linguistic features are more difficult in all three languages. We see this as a reasonable assumption, especially since all languages in the study are Germanic languages and have much in common. Still, it is possible that the relation between linguistic features and difficulty is different in the different languages (Bergqvist et al., 2018). Our method of analysis was not designed to detect such varying relationships. More detailed analyses on groups of tasks would be necessary to enable detection of these kinds of relationships.

### *Conclusion*

In this study, mathematics tasks displaying DIF between different language versions were identified in the PISA 2012 assessment. There must be a reason for this functional inequivalence, for example, other linguistic differences than what we examined. It is also possible that cultural or curricular differences between the language groups are reasons for DIF. Furthermore, for a specific task, there can be several different reasons for DIF, which could interact with each other. Such interaction could hide any effects from one specific reason, which can be another explanation for our results showing no statistical relations. More research is needed to identify reasons for DIF and potential interactions between different reasons. This can be done with quantitative methods, as in the current study, or using qualitative methods, such as expert reviews of the tasks or methods including students' experiences from different language groups who are working with the tasks. By identifying reasons for DIF, it may be possible to make changes in the tasks or replace tasks to minimize the occurrence of DIF. By that, a higher degree of equivalence of the tasks' language versions is obtained and the validity of the results of multilanguage assessments can be enhanced.

In the translation and adaptation guidelines for PISA 2012 (OECD, 2010, p. 12) it is recommended to "avoid, translating an active turn of

phrase in the original by a passive one, or vice versa" if possible, since passive voice is regarded as being more difficult. Despite this recommendation, it happened several times in the tasks investigated in this study that different voice was used in the different language versions. However, if functional equivalence is prioritized, which is often the case in assessment, our results suggest that differences between language versions concerning voice, grammatical person and sentence structure are not that important to focus on. The present study also points to the importance of further studies of the effects of the translations in multilanguage assessments, since the relation between functional and formal equivalence regarding these linguistic features is not yet clarified. If further studies would support the results in our study, that differences in the use of voice, and also grammatical person and sentence structure, are not related to functional equivalence of tasks, the recommendation in the PISA guidelines could either be removed completely or at least be given a lower level of importance. Translators could then instead focus on other features that might affect functional equivalence between the language versions more.

## References

Abedi, J., Lord, C. & Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance* (CSE Report 429). National Center for Research on Evaluation, Standards, and Student Testing.

AERA (2014). *Standards for educational and psychological testing.* American Educational Research Association.

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

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in Education*, 16 (1), 55–73.

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.

Allan, S. (2009). *Passive be damned: the construction that wouldn't be beaten* (Master of Arts). University of Canterbury,

Allen, N. L. & Donoghue, J. R. (1996). Applying the Mantel-Haenszel procedure to complex samples of items. *Journal of educational measurement*, 33 (2), 231–251.

Arffman, I. (2010). Equivalence of translations in international reading literacy studies. *Scandinavian Journal of Educational Research*, 54 (1), 37–59.

Bergqvist, E., Theens, F. & Österholm, M. (2018). The role of linguistic features when reading and solving mathematics tasks in different languages. *The Journal of Mathematical Behavior*, 51, 41–55.

Berry, R. (2012). *English grammar: a resource book for students*. Routledge.

Bostian, L. R. (1983). How active, passive and nominal styles affect readability of science writing. *Journalism Quarterly*, 60 (4), 635–640, 670.

*Chicago manual of style* (2003) (15 ed.). University of Chicago Press.

Çikrikçi Demirtaşli, N. & Ulutaş, S. (2015). A study on detecting of differential item functioning of PISA 2006 science literacy items in Turkish and American samples. *Eurasian Journal of Educational Research*, 58, 41–60.

Davis-Dorsey, J., Ross, S. M. & Morrison, G. R. (1991). The role of rewording and context personalization in the solving of mathematical word problems. *Journal of Educational Psychology*, 83(1), 61.

Dorans, N. J. & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization. *ETS Research Report Series*, 1992 (1), i–40.

Dryer, M. S. & Haspelmath, M. (2013). *The world atlas of language structures online*. http://wals.info

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), 23–35.

Ferne, T. & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, 4 (2), 113–148. doi: 10.1080/15434300701375923

Forster, K. I. & Olbrei, I. (1973). Semantic heuristics and syntactic analysis. *Cognition*, 2 (3), 319–347.

Grisay, A. & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation*, 33 (1), 69–86.

Hickendorff, M. (2013). The language factor in elementary mathematics assessments: computational skills and applied problem solving in a multidimensional IRT framework. *Applied Measurement in Education*, 26 (4), 253–278. doi: 10.1080/08957347.2013.824451

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J. & Baird, J.-A. (2017). Lessons learned from PISA: a systematic review of peer-reviewed articles on the programme for international student assessment. *Scandinavian Journal of Educational Research*, 1–21.

Huang, X., Wilson, M. & Wang, L. (2014). Exploring plausible causes of differential item functioning in the PISA science assessment: language, curriculum or culture. *Educational Psychology* (ahead-of-print), 1–13.

Koller, W. (2011). *Einführung in die Übersetzungswissenschaft* (8., neubearb. Aufl. ed.). Francke.

Lord, C. (2002). Are subordinate clauses more difficult? In J. L. Bybee & M. Noonan (Eds.), *Complex sentences in grammar and discourse: essays in honor of Sandra A. Thompson* (pp. 223–234). John Benjamins.

Michaelides, M. P. (2008). An illustration of a Mantel-Haenszel procedure to flag misbehaving common items in test equating. *Practical Assessment Research & Evaluation*, 13 (7).

O'Grady, K., Karen, F., Servage, L. & Khan, G. (2018). *PCAP 2016. Report on the Pan-Canadian assessment of reading, mathematics, and science.* Council of Ministers of Education. https://cmec.ca/507/PCAP_2016.html

OECD (2010). *Translation and adaption guidelines for PISA 2012*. OECD.

OECD (2012). *Data base – PISA 2012*. OECD. http://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: a comparison of three Mantel–Haenszel procedures. *Applied Measurement in Education*, 14 (3), 235–259.

Pym, A. (2010). *Exploring translation theories*. Routledge.

Ross, S. M. & Anand, P. G. (1987). A computer-based strategy for personalizing verbal problems in teaching mathematics. *ECTJ*, 35 (3), 151–162.

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.

Solano-Flores, G., Backhoff, E. & Contreras-Niño, L. Á. (2009). Theory of test translation error. *International Journal of Testing*, 9 (2), 78–91.

Swedish National Agency for Education (2018). *Beställning av nationella prov för höstterminen 2018*. Skolverket. www.skolverket.se/download/18.4fc05a3f164131a7418259/1533891917368/bestallningsbrev-np-ht-gymnasieskolan-2018-2019.pdf

Taylor, A. B., West, S. G. & Aiken, L. S. (2006). Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. *Educational and Psychological Measurement*, 66 (2), 228–239.

Theens, F. (2019). *Does language matter? Sources of inequivalence and demand of reading ability of mathematics tasks in different languages* (PhD thesis). Umeå universitet.

Villiers, J. G. de & Villiers, P. A. de (1973). Development of the use of word order in comprehension. *Journal of Psycholinguistic Research*, 2 (4), 331–341.
Wichmann, S., Holman, E. W. & Brown, C. H. (2016). *The ASJP database* (version 17). http://asjp.clld.org
Volansky, V., Ordan, N. & Wintner, S. (2013). On the features of translationese. *Digital Scholarship in the Humanities*, 30 (1), 98–118.
Xiao, R. & Yue, M. (2009). Using corpora in translation studies: the state of the art. In P. Baker (Ed.), *Contemporary corpus linguistics* (pp. 237–261). Continuum.
Yildirim, H. H. & Berberoğlu, G. (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *International Journal of Testing*, 9 (2), 108–121.
Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Directorate of Human Resources Research and Evaluation.

## *Notes*

1 Some PISA items are polytomous with scores 0, 1, or 2. For the DIF-analysis, these items were dichotomized by setting scores 1 and 2 as full credit and 0 as no credit (see e.g. Michaelides, 2008; Yildirim & Berberoğlu, 2009).

2 The Mantel-Haenszel method results in a measure called MH D-DIF. The categorization in the three levels A, B and C depends both on how big the absolute value of this measure is and how significantly it exceeds a certain value. If the MF D-DIF measure for a particular test is not significantly different from zero or less than 1.0 in absolute value, the level of DIF is classified as A (negligible). If the MH D-DIF value differs significantly from zero and is greater than 1.0 in absolute value but is not significantly greater than 1.0 or smaller than 1.5 in absolute value, the level of DIF is classified as B (intermediate). If the MH D-DIF value is significantly greater than 1.0 and greater than 1.5 in absolute value, the level of DIF is classified as level C (large) (Dorans & Holland, 1992). This categorization is commonly used (see e.g. Allan & Donoghue, 1996; Michaelides, 2008).

3 Also imperative mood was counted as second person, since "although there is no subject and auxiliary with imperatives, it can be suggested that the underlying subject is you [...]" (Berry, 2012, pp. 123–124). Since first person ("I", "we") was not used at all in the mathematics PISA tasks, it was not included in the calculations.

## Appendix

In table A1, the tasks displaying DIF between the different language versions are listed for each pair of language versions. Empty cells imply negligible DIF (level A) between the language versions. The numbers of the tasks are the ones used in official PISA documents.

*Table A1. Tasks displaying DIF between the language versions*

| Task number | SWE-GER | | ENG-GER | | ENG-SWE | |
|---|---|---|---|---|---|---|
| | Favored language version | DIF-level | Favored language version | DIF-level | Favored language version | DIF-level |
| PM155Q04 | SWE | B | | | | |
| PM192Q01 | | | | | SWE | C |
| PM305Q01 | | | GER | B | SWE | C |
| PM406Q01 | SWE | B | | | SWE | C |
| PM408Q01 | | | GER | B | SWE | B |
| PM420Q01 | | | ENG | C | | |
| PM442Q02 | | | | | ENG | B |
| PM446Q01 | SWE | B | ENG | C | ENG | C |
| PM447Q01 | GER | B | | | | |
| PM464Q01 | | | GER | C | SWE | B |
| PM474Q01 | | | | | SWE | B |
| PM559Q01 | SWE | B | | | SWE | B |
| PM603Q01 | | | GER | B | | |
| PM800Q01 | SWE | B | | | SWE | C |
| PM828Q01 | SWE | B | | | | |
| PM828Q03 | | | GER | B | | |
| PM903Q01 | | | | | ENG | B |
| PM903Q03 | | | ENG | C | ENG | C |
| PM905Q02 | | | GER | B | SWE | B |
| PM906Q01 | | | GER | B | SWE | B |
| PM909Q01 | | | | | ENG | B |
| PM909Q03 | GER | B | GER | B | | |
| PM915Q01 | SWE | C | ENG | B | | |
| PM915Q02 | GER | C | GER | C | | |
| PM918Q01 | | | ENG | B | ENG | B |
| PM919Q01 | | | GER | B | SWE | B |
| PM923Q01 | | | | | SWE | B |
| PM949Q02 | GER | B | ENG | B | ENG | C |
| PM949Q03 | | | ENG | B | ENG | B |
| PM953Q02 | | | ENG | B | | |
| PM953Q03 | | | | | SWE | B |
| PM953Q04 | | | ENG | B | | |
| PM954Q01 | | | ENG | B | ENG | C |
| PM955Q02 | | | GER | B | SWE | B |
| PM982Q02 | GER | B | ENG | B | ENG | C |
| PM992Q02 | SWE | B | | | | |
| PM995Q01 | GER | B | | | | |
| PM998Q02 | | | ENG | B | | |
| PM998Q04 | | | | | SWE | B |
| PM00GQ01 | | | | | SWE | B |
| PM00KQ02 | | | GER | C | SWE | C |

## Frithjof Theens

Frithjof Theens is a senior lecturer at the Department of Science and Mathematics Education at Umeå University and has a PhD in mathematics education. He is a member of Umeå Mathematics Education Research Centre (UMERC). His main research interest is in language connected to mathematics education.

frithjof.theens@umu.se

## Ewa Bergqvist

Ewa Bergqvist is a docent (associate professor) in mathematics education at the Department of Science and Mathematics Education at Umeå University. She is a member of Umeå Mathematics Education Research Centre (UMERC) and leads a research group focusing on language and communication in mathematics together with Magnus Österholm. Her research focuses mainly on aspects of argumentation, language, and curriculum implementation in school and university level mathematics.

ewa.bergqvist@umu.se

## Magnus Österholm

Magnus Österholm is a professor of mathematics education at Umeå University and at Mid Sweden University. He leads a research group focusing on language and communication in mathematics together with Ewa Bergqvist. His main research interest is the role of language and communication in the learning and teaching of mathematics, but he has also studied issues in relation to belief-research and teachers' professional development.

magnus.osterholm@umu.se