

The Trade-Off between Model Fit, Invariance, and Validity: The Case of PISA Science
Assessments

Abstract

In large-scale educational assessments, it is generally required that tests are composed of items that function invariantly across the groups to be compared. Despite efforts to ensure invariance in the item construction phase, for a range of reasons (including the security of items) it is often necessary to account for differential item functioning (DIF) of items *post hoc*. This typically requires a choice among retaining an item as it is despite its DIF, deleting the item, or resolving (splitting) an item by creating a distinct item for each group. These options involve a trade-off between model fit and the invariance of item parameters, and each option could be valid depending on whether or not the source of DIF is relevant or irrelevant to the variable being assessed. We argue that making a choice requires a careful analysis of statistical DIF and its substantive source. We illustrate our argument by analyzing PISA 2006 science data of three countries (UK, France and Jordan) using the Rasch model, which was the model used for the analyses of all PISA 2006 data. We identify items with real DIF across countries and examine the implications for model fit, invariance, and the validity of cross-country comparisons when these items are either eliminated, resolved or retained.

Keywords: Rasch model, differential item functioning, model fit, validity, invariance, cross-country comparison, international large-scale assessments

The Trade-Off between Model Fit, Invariance, and Validity: The Case of PISA Science Assessments

The quality of an educational assessment rests on the degree to which students' test scores reflect their knowledge, understanding, and skills on the construct being assessed, and the degree to which they are consistent under similar testing conditions. The former is referred to as the test's *validity*; the latter as its *reliability*. In this paper we are concerned with quantitative comparisons and therefore refer to the quantified form of the assessment using a test's scores as a *variable*.

Providing the items are functioning as required, both validity and reliability increase with the number of items. Validity increases because the number of different, generally distinct, *aspects* of the same variable that can be assessed increases. Reliability increases because the number of distinct classifications on the variable increases. In addition, the quality of an assessment depends on its capacity to offer all students equal opportunities to demonstrate their knowledge, understanding, and skills in regard to the variable. In educational contexts, this may be referred to as students' *proficiency*. First, the range of item difficulties should permit students with a range of proficiencies to demonstrate their standing on a variable. Second, proficiency should be assessed with the same validity and reliability irrespective of the students' group membership (for example, country, gender, ethnicity, and socioeconomic status).

In general terms, assessments are expected not to disadvantage a particular group by having features of the items that are *irrelevant* to the variable being assessed. When the performance of equally proficient students on an item is not invariant with respect to group membership, and the lack of invariance is attributed to a feature of that item that is not relevant to the variable being assessed, then in everyday language, the item is said to be

biased. Bias against groups is a threat to the validity of an assessment and to any comparisons made among groups.

Employing a common, descriptive, and technical term (Holland & Thayer, 1988), items that do not function invariantly across groups are said to exhibit differential item functioning (DIF). The conceptualization for understanding DIF that is most relevant to the data analyzed in this paper, where the items are designed to assess not only the main variable of assessment, but three distinct *aspects* of this variable, is that it is a form of multidimensionality (Ackerman, 1992). The variable and its three distinct aspects are presented in a subsequent section. The presence of DIF in these data is taken to indicate that the assessment of the distinct aspects functions differently across groups.

The distinct aspect of an item that functions differently could have been included deliberately or unintentionally by item writers where generally the former may be relevant to the variable and the latter irrelevant. We now give an example of each type (deliberate and intentional), with their implications. First, if the distinct aspect and source of DIF in an item is relevant to the assessed variable (e.g., the technical term *chlorophyll* in an item assessing students' understanding of photosynthesis), then the item is considered sound and the DIF reflects group differences that are justifiable (e.g., one group being taught technical vocabulary more effectively than other groups). Second, if the distinct aspect and source of DIF in an item is irrelevant to the assessed variable, (e.g., an ambiguous instruction in assessing students' understanding of photosynthesis), then the item is not considered sound and the DIF reflects group differences which are unjustifiable.

Because DIF in some items can lead to unjustifiable differences in the comparison of groups, methods for identifying DIF have been developed (Penfield & Camilli, 2007; Zumbo, 2007). These methods are facilitated by models of modern test theory in which the response

of a person to an item is characterized by a probabilistic model (Birnbaum, 1968; Van der Linden, 2016). One such model is the Rasch model (Andersen, 1977; Andrich, 1988; Fischer & Molenaar, 1995; Rasch, 1960), which has been used in large-scale assessment studies and is also used in this paper. The model itself is characterized by the property of invariance (Rasch, 1960; Duncan, 1984), and hence lends itself to the study of the trade-off between invariance and model fit, and the extent to which the trade-off may affect the validity of comparisons among groups.

Ideally, items that exhibit DIF can be rewritten to eliminate their DIF and improve their fit while simultaneously retaining the validity and reliability they contribute. However, in many cases, especially in the context of high-stakes assessments, rewriting items is not practical – pretesting items before their live administration is time consuming, expensive, and weakens the security of the items. In addition, even when items are pretested, because DIF could be a manifestation of an interaction between different factors distinguishing groups (Penfield & Camilli, 2007), DIF could still arise in the assessment. Indeed, data analysis of international large-scale assessments such as the Programme for International Student Assessment (PISA), which we illustrate in this study, provides empirical evidence of DIF across national versions of the same test (Asil & Brown, 2016; Ercikan & Koh, 2005; Grisay, de Jong, Gebhardt, Berezner, & Halleux-Monseur, 2007; Grisay, Gonzalez, & Monseur, 2009; Hauger & Sireci, 2008; Huang, Wilson, & Wang, 2016; Kreiner & Christensen, 2014; Le, 2009; Oliveri, Olson, Ercikan, & Zumbo, 2012; Sandilands, Oliveri, Zumbo, & Ercikan, 2013; Wu & Ercikan, 2006; Xie & Wilson, 2008). It is then common to resort to statistical methods for the *post hoc* accounting of DIF.

The two most commonly adopted statistical methods for handling DIF *post hoc* (besides retaining an item as it is despite its DIF) are: (1) *deleting*, and (2) *resolving* the item

(Andrich & Hagquist, 2012; Kreiner & Christensen, 2014). Deleting an item with DIF results in improved fit of the responses to the model. However, in deleting an item, the distinct aspect of the variable assessed by the item is removed from the test. If this distinct aspect is relevant to the variable, then its removal reduces the validity of the assessment. Moreover, because each item also assesses the main variable, deleting an item typically reduces the reliability and the precision of estimates of proficiency.

The Resolution of Items

Therefore, the preferred approach to dealing statistically with the DIF of an item is to resolve it. We now elaborate this procedure and its features. As indicated above, resolving an item involves creating a distinct item for the responses of each group. Thus if there are three groups, it is as if there were three items, but each one answered by only one group. This approach creates structural missing data that modern software handles routinely. In this paper, we use the term *resolving*, the terminology used in Andrich and Hagquist (2012, 2015) and Hagquist and Andrich (2017), instead of the alternative term *splitting*. This choice helps emphasize that the resolution of a DIF item is in terms of its *constituent* components that in this case are the groups (e.g., gender, language, socioeconomic status, country, etc.) that make up the larger sample.

Resolving, rather than deleting, an item has the following advantages. First, because the item is retained, an assessment of the main aspect of the variable is retained. Second, because each group still answers the same number of items, the precision afforded by the item is also retained. Third, as with deleting an item, and as illustrated with the example below, the relevant statistics will show improved model fit.

Fourth, a more subtle but no less important property is that it bears directly on the distinction between *real* and *artificial* DIF (Andrich & Hagquist, 2012, 2015). Andrich and

Hagquist explain artificial DIF as an *artefact* of typical methods of detecting DIF in which real DIF in an item favoring one group induces DIF in the other items favoring other groups. This induced artificial DIF appears because the estimates are used to classify students by proficiency in determining whether students of the same proficiency estimate, but from different groups, perform differently on an item. In the Rasch model, the estimate of proficiency is based on a person's total score. Then, given the same total score, if a group has a higher proficiency on an item because of real DIF, the group will have a relatively lower score on other items. This artificial DIF is present in many methods of detecting DIF, including the popular (and seen as a standard) Mantel-Haenszel method (Holland, & Thayer, 1988). Andrich and Hagquist (2012, 2015) also explain how, when an item with real DIF is resolved, then it no longer induces artificial DIF in the other items.

It is stressed that the concern here is with real DIF, and therefore, in the analysis of the impact of dealing with DIF statistically, it is relevant to remove artificial DIF. As a result, to identify real DIF, items are resolved sequentially beginning with the one that has evidence of the greatest DIF.

Fifth, a resolved item for each group provides a different estimate of its difficulty for each group, and the different difficulty estimates immediately quantify the effect of DIF in the metric of the proficiency estimates. To make the point, if an item with no real DIF is resolved, then the difficulty estimates from the different groups will be statistically equivalent.

Sixth, in the estimate of person proficiencies, these differences in item difficulties immediately compensate for DIF. Thus, if a group finds an item relatively more difficult, then answering the item correctly rewards that group more than it rewards the group that finds the item relatively easier.

Although there are these advantages in resolving an item, there is one major disadvantage. This is that in resolving an item, its relative difficulty is *not invariant* among the groups – different groups have a unique difficulty for the item. Using data from PISA science assessment from three countries, we make tangible the issue of whether or not this factor detracts from, or improves, the validity of comparisons.

We analyzed a subset of PISA 2006 data using the Rasch model that was used also with the complete PISA data. The magnitude of observable DIF was identified using a two-way ANOVA of residuals, in which one factor was membership in class intervals on the variable, and the other was membership of the country (Andrich & Hagquist, 2012, 2015). This method of determining the magnitude of DIF has the advantage that it immediately characterizes both uniform and non-uniform DIF. Uniform DIF occurs when differences between means among different countries are statistically equivalent across class intervals, while non-uniform DIF is manifested when the differences between means among different countries across class intervals are not consistent. Then, with a view to identifying and quantifying real DIF and removing any artificial DIF induced in other items, the item with the greatest DIF was resolved, and the data reanalyzed. The same procedure was repeated with the item that showed the greatest magnitude of DIF in the second analysis. The procedure was repeated until it was considered that there was no further substantive evidence of real DIF.

In summary, using the PISA dataset illustratively, and following the procedure summarized above, we argue, that analyses and interpretations of educational assessment data inevitably involve a degree of trade-off between model fit and invariance of item parameters. Consequently, the decision to resolve an item and improve fit or not to resolve the item and

retain the invariance, should be guided by an understanding of the source of DIF: this understanding is external to the analysis.

The Variable and the Dataset

In this section, we summarize the complex nature of the scientific literacy variable measured by PISA science assessments, describe the dataset selected in this study, and elaborate its analysis based on the general principles described above.

The Variable of Scientific Literacy in PISA: A Latent and a Composite Variable

PISA is an international survey administered by the Organization for Economic Cooperation and Development (OECD) every three years since 2000 in over 70 countries. The survey assesses 15-year-olds' literacy in reading, mathematics, and science with other domains being optional (e.g., financial literacy). In each cycle, one of the compulsory domains is assessed as a major domain while the other two are assessed as minor domains. The 2006 cycle, from which data for this study has been retrieved, was the first of two instances in which science literacy was the major domain; the second being in 2015. The source versions of PISA are developed in English and French, except for a subset of items proposed by participating countries, and translated following a thorough double translation approach into the target languages (OECD, 2009). In 2006, PISA was translated into over 40 languages.

It is assumed that students' literacy in science, i.e., their knowledge and skills in science (OECD, 2006, 2017), underlies their performance on PISA science assessments. In this sense, *scientific literacy* can be conceptualized as consisting of a variable that *causes* performance on a test. This belief is in keeping with Edwards and Bagozzi's (2000) conceptualization of reflective relationships where the causal relationship holds in the

direction that students' literacy causes their performance on science items (rather than the opposite direction). One of the implications of this conceptualization is that the items of a science test become, *in principle*, exchangeable. Hence, deleting or resolving an item from a science test would not technically affect a student's estimate of proficiency, other than its effect on precision. In practice, however, even in a fully causal context, items tend not to be fully exchangeable; for example, very easy and very difficult items are typically not exchangeable for the range of science proficiencies to be measured.

The scientific literacy adopted in PISA assessments can, however, also be considered as a *composite* variable that consists of a number of discrete yet correlated indicators. These have also been described as formative models (Edwards & Bagozzi, 2000). The theoretical variable is then the *effect* of the indicators combined. Stenner, Burdick, and Stone (2008) illustrated such a variable with the concept of socioeconomic status (SES), which they defined as the level of education, occupational prestige, level of income, and the desirability of the neighborhood in which people live. They stress that although these indicators are correlated and therefore justify a single index of SES, the indicators together define SES rather than the SES causing a response on these indicators. As a consequence, the items in principle are not exchangeable. For example, *neighborhood* is not exchangeable for *income* because if the indicator of *income* were eliminated from the index, then in a real sense the variable would have been altered.

Analogously, scientific literacy has been defined for PISA 2006 as:

... the capacity to use scientific knowledge, to identify questions and to draw evidence-based conclusions in order to understand and help make decisions about the natural world and the changes made to it through human activity.

Source: OECD, 1999, p. 60¹

The variable has been operationalized as consisting of three distinct aspects described as competencies: (1) identifying scientific issues, (2) explaining phenomena scientifically, and (3) using scientific evidence (OECD, 2006). Each competency is assessed by at least 25% of the items, with the first competency (i.e., identifying scientific issues) counting for fewer points (25-30 points) than the other two competencies, each of which count for 35 and 40 points, respectively (OECD, 2006). In more general terms, scientific literacy as defined for PISA 2006 science assessments is *composed* of three distinct aspects, and each PISA item focuses on at least one of these. This composition and their relationship to the items of assessments are summarized in Figure 1.

<<INSERT FIGURE 1 ABOUT HERE>>

A consequence of conceptualizing scientific literacy as a composite variable in which each item assesses a distinct as well as the common aspect of the variable, is that the items are not, in principle, exchangeable. By analogy to the example of income and neighborhood

¹ Note that the construct of scientific literacy has evolved to “the ability to engage with science-related issues, and with the ideas of science as a reflective citizen” in PISA 2015 (OECD, 2017).

not being exchangeable in the Stenner et al. (2008) SES example, in PISA science 2006, items assessing proficiency in identifying scientific issues are not exchangeable with those assessing proficiency in explaining phenomena scientifically or with ones assessing proficiency in using scientific evidence properly.

Scientific literacy here mirrors Andrich's (2014) example of the variable of proficiency in physics that exhibits features of both a causal and a composite variable. Andrich explained how proficiency in physics could be viewed as weaving together (much like the strands of a rope) proficiencies in five sub-disciplines (heat, light, sound, electricity and magnetism, and mechanics) and used the terms *thick* for the composite variable of physics and *thin* for each of the sub-disciplines. Then at the level of a thin variable, the items can be considered causal, while at the level of the thick variable, they need to be considered as composite. Tesio (2014) showed that such a conceptualization was also relevant to health outcomes assessment.

In addition to taking PISA's definition of scientific literacy to make it both a causal and a composite variable, by referring to the science education literature (e.g., Norris & Phillips, 2003) and using a critical evaluation of released PISA 2006 science items, El Masri, Baird and Graesser (2016) argued that language is an inextricable part of the construct of scientific literacy. Referring to Figure 1, if for the sake of the argument, one is to conceive of the three science competencies of the scientific literacy variable as common across countries and hence comparable, El Masri et al. (2016) argued that the language component is inevitably unique for each country. The authors provided examples such as the translation of the phases of the moon cycle from English (e.g., crescent) to French (*croissant*) or Arabic (*[al-hilal]* الهلال) leading to terms with different word frequencies across languages and hence placing different cognitive demands on students across language groups. For example, *croissant* in French refers to the shape of the famous French pastry and the equivalent in Arabic refers to the

symbol of Islam with which Arabic-speaking students are most likely familiar. The language aspect inevitably adds to the complexity of the scientific literacy construct adopted in PISA.

The Dataset: PISA 2006 Science

The dataset analyzed and reported in this paper is that of PISA 2006 science assessments of three countries, the United Kingdom (UK), France, and Jordan. The assessments consist of 103 cognitive science items² (33 dichotomous, six polytomous with three response categories, and 64 multiple choice items) administered in the PISA 2006 cycle to 15-year-olds in the UK, France, and Jordan in each country's language of instruction; that is, English, French, and Arabic, respectively. The overall sample across the three selected countries consisted of 24 377 students.

The purpose of having a large set of 103 items was to cover the range of science proficiencies conceptualized as a composite variable as in Figure 1. However, not all students could be allocated time to answer all 103 items. Therefore, the 103 items were administered in a balanced, incomplete block design where clusters were rotated in a specific way to form 13 test booklets with overlapping items which permits placing their difficulty estimates on the same scale (OECD, 2009, p. 39).

As shown in Table 1, the sample sizes among the selected countries, were noticeably different. Therefore, to eliminate any weighting by country in the DIF analysis, for the purposes of this paper a random sample of 4 000 students from each country was chosen. In addition, we note that because the power of detecting misfit is a function of sample size and, because no data can fit any model perfectly, a very large sample will have such power that every item will show substantial misfit. The number of students who answered each item

² Non-cognitive items assessing students' attitudes and interest in science were not considered in this study.

ranged from 3 489 to 3 672 - a substantial number and more than powerful enough to detect any misfit.

<<INSERT TABLE 1 ABOUT HERE>>

Data Analysis

The Rasch model for dichotomous and polytomous items (Andrich, 1978; Rasch, 1960) was used to estimate item and person parameters of PISA 2006 science data of the three countries. The analyses were carried out using the RUMM2030 software (Andrich, Sheridan, & Luo, 2015) that uses the conditional pairwise estimation algorithm for estimating the item parameters independently of the person parameters and therefore independently of any distribution of the persons. Given the item difficulties, the person proficiencies are then estimated individually, based on the items to which each person was asked to respond. The estimates used are a weighted modification of maximum likelihood estimation, a weighting that reduces bias in the estimates of proficiencies of scores, which are near the maximum or minimum scores on a test (Warm, 1989).

Results of Model Fit and DIF Analysis

In the analysis of fit reported below, four class intervals of students were formed across the continuum of item/person locations. In the interpretation of fit, some preliminary points are emphasized. First, no available fit statistic is both necessary and sufficient for detecting all possible violations of a model that can be present in the data; different fit statistics focus on detecting different violations. Second, fit is relative, not absolute, with its being a matter

of degree, and with the degree of fit required being a function of the substantive context. Third, the power of a test of fit is a function of the alignment between the persons and items and the sample size. Thus, floor and ceiling effects create problems of bias in the interpretation of estimates, and the sample sizes used for survey purposes can be so large that they show that all items misfit. Fourth, unlike many tests of hypotheses, the tests of fit are deviations from perfection, not deviations from randomness.

Finally, no item can exhibit fit on its own, and generally more than two items are required to be able to conduct any test of fit. Instead, *misfit of an item* is shorthand for evidence that the item does not work consistently with other items in providing a scalar estimate on a single variable as summarized by the model used for the analysis: here, the Rasch model. The item characteristic curve (ICC) of the expected value of a response as a function of the value on the proficiency variable, provides the frame of reference for the detection of DIF and overall fit, irrespective of the group. In the case of dichotomous items, the expected value curve is also the probability of a correct response. Such curves are shown with examples in the next section. Items with substantial deviation from this curve are taken to misfit, with the particular fit emphasized in this paper being DIF.

Person – Item Threshold Distribution

Before proceeding with the detailed results of the item analysis focusing on DIF, it is helpful to examine the extent to which the distribution of the countries' proficiencies and the item difficulties are aligned before any of the items were resolved. Figure 2 shows that there are some students with proficiency estimates below the easiest item on the scale and also some above the most difficult items. However, for the purpose of the analysis of DIF, Figure 2 indicates that the item difficulties are well aligned to the student distribution.

<<INSERT FIGURE 2 ABOUT HERE>>

DIF Analysis and Fit of DIF Items

DIF across countries has been detected in other international large-scale studies (Asil & Brown, 2016; Ercikan & Koh, 2005; Grisay et al., 2007; Grisay et al., 2009; Hauger & Sireci, 2008; Huang et al., 2016; Kreiner & Christensen, 2014; Le, 2009; Oliveri et al., 2012; Sandilands et al., 2013; Wu & Ercikan, 2006; Xie & Wilson, 2008). For the purposes of this paper, and because of the large number of items in the dataset, we present the results of the DIF analysis only for those items deemed to have real DIF as identified by the sequential procedure described above.

There were 23 such items³. These items, with the ANOVA residuals, are shown in Table 2. For completeness, the table also shows the fit across the class intervals, which is a general test of fit. Because none of these items exhibited non-uniform DIF, Table 2 does not show the statistics for the interaction between the class intervals and the countries. No non-uniform DIF implies that the 23 items behaved consistently differently in at least two countries across all class intervals.

<<INSERT TABLE 2 ABOUT HERE>>

³ Due to the large number of comparisons in this analysis, Bonferroni correction (Dunn, 1961) was applied to adjust the significance level to $p < 0.000485$ to reduce the risk of familywise error rate.

Table 2 shows that two of these 23 items (items 1 and 96) not only showed DIF, but also misfit across the class intervals. To illustrate graphically the information in Table 2, Figure 3a shows the plot of the means of the class intervals relative to the ICC for item 1 (a dichotomous item), without (left) and with (right), the classification by country. It is evident that, irrespective of country, the item discriminates too highly relative to the average discrimination of all items as reflected in the common slope of the ICC of all items. It is also evident that, for the same estimate of proficiencies based on all items answered by the students, students in Jordan do not perform as well as those from the other two countries.

<<INSERT FIGURE 3a ABOUT HERE>>

To illustrate DIF in an item that otherwise fits across the class intervals (which is the case for the majority of items in Table 2), Figure 3b shows graphically the plot of the means of the class intervals relative to the ICC for Item 43 (a dichotomous item), without (left) and with (right), the classification by country. Unlike Item 1, the means of the class intervals, when countries are not identified, are close to the ICC, and again unlike Item 1, the students in Jordan, for the same estimate of proficiency, perform relatively better on this item.

<<INSERT FIGURE 3b ABOUT HERE>>

Four points are noted from Table 2 and Figures 3a and 3b. First, 22% of items in the dataset were taken to exhibit real DIF. Second, the two tests of fit used in this analysis, the

general test of fit and DIF, deal with different violations of the model and that to identify a specific kind of misfit, it is necessary to use the relevant statistic. Third, the overall performance of a country does not preclude DIF, which shows the country may overperform or underperform on the item relative to the other countries. Fourth, although there may be misfit, irrespective of country, the general trend of the performance in the class intervals follows the ICC in that the means of the successive class intervals increase.

Implications of Resolving DIF Items

Given that the proportion of DIF is substantial (22%), resolving items with real DIF in this dataset is likely to have implications for overall model fit, invariance, and consequently the validity of comparisons. We examine these in this section beginning with an examination of the effects on model fit.

Implications of Resolving Items for Model Fit

Because of the nested nature of the analyses before and after resolving items, it is possible to quantify the magnitude of the improvement in fit. Given estimates of the item and person parameters, the log likelihood of the data can be calculated from each analysis. With 46 more parameters estimated from the resolution of the DIF items (i.e., 2 by 23), it is expected that the likelihood of the data, given the parameter estimates, would be greater than in the original analysis. The significance of the difference can be assessed with a log likelihood ratio test with the relevant statistic distributed under the hypothesis of no difference (asymptotically) as the chi square distribution, with degrees of freedom being the difference in the number of parameters from the two analyses (i.e., 46). Table 3 shows the likelihoods from the two analyses, the resultant chi square statistic, and the degrees of freedom of such a statistic under the null hypothesis of no difference. With the expected

value of a chi square being its degrees of freedom, in this case 46, the chi square statistic of 5157.07 is hugely significant. Thus, there is a substantial improvement in overall fit following the resolution of the 23 items. From the sole perspective of model fit, the improved fit would be paramount.

<<INSERT TABLE 3 ABOUT HERE>>

However, although fit can be improved by resolving the items showing DIF, we have emphasized throughout that the difficulties of these items are no longer invariant across countries. Having a substantial number of items with different difficulty estimates, not only between countries, but different from the original estimates before resolution, can lead to noticeable differences in proficiency estimates. This difference between analyses translates into implications for the validity of country comparisons. We investigate more closely the lack of invariance of item difficulties across countries between the original and the resolved analyses, and the impact of this change on person proficiency estimates.

Implications for Item Difficulty Estimates

Table 4 shows the difficulty estimates from the original analysis and for each country from the resolved analysis. The mean of the difficulties of the resolved items for each country reflects the impact of DIF at the test level. Table 4 also shows these mean difficulty estimates by country.

<<INSERT TABLE 4 ABOUT HERE>>

Table 4 shows that, overall, the mean of the resolved DIF items is easiest for Jordan (-0.225) and most difficult for France (0.186). However, this variation is not uniform among items, with some more difficult for one country and others more difficult for other countries. It is relevant to note that, as shown in Figure 2, Jordan - which had the lowest overall proficiency in science - had the easiest mean of the resolved DIF items, conditional on equivalent proficiencies. The reason for this effect cannot be explained from the analysis alone, and the source of DIF that led to this outcome must be obtained from external sources. Focusing on a possible language source, it may be that these items are more readily and better translated into Arabic than the other items. Therefore, in this specific case, it reinforces Grisay and colleagues' (Grisay et al., 2007; Grisay & Monseur, 2007) call for further research into the comparability of Arabic versions of international tests with the source versions (English and French, for PISA).

To test whether the means of the locations of resolved items were significantly different among countries, a repeated measures ANOVA was carried out⁴. The details of this ANOVA are displayed in Table 5.

<<INSERT TABLE 5 ABOUT HERE>>

The condition of sphericity of variances was met (Mauchly's test statistic = 0.92; probability = 0.39 > $p = 0.05$). Results shown in Table 5 indicate that the within-subject effects were not significant ($p = 0.130 > 0.05$), suggesting that there was no significant net

⁴ A repeated-measures ANOVA was considered rather than a one-way ANOVA because there is dependence among the country-specific item locations in every row; they all relate to the same item.

advantage or disadvantage given to any country by DIF items. For a given country, the advantages conferred by specific items were compensated for by disadvantages that were inherent in others. This finding is consistent with similar studies carried out on PISA data where, for instance, Oliveri et al. (2012) found that the English and French versions of the PISA 2003 problem-solving test in Canada showed a high degree of comparability at test level, and a lower degree at item level. Hence, in this particular case, although invariance was not maintained at the item level after resolving DIF items, it was maintained at the test level because real DIF among items within each country balanced out. Although the means of the DIF items were not significantly different among the countries, these differences will nevertheless have some impact on the means of the proficiencies of the groups relative to the original analysis. For illustrative purposes, this impact is reported below.

Implications for Person Estimates

The direction of the change in proficiency estimates before and after resolving items can be anticipated from the average item difficulties in Table 4. Because the mean of the resolved item difficulties is easiest for Jordan and most difficult for France, the students in Jordan will receive relatively less reward for a correct response to these items than will students from France (Andrich & Hagquist, 2012). Therefore, it can be anticipated that to the degree that there is any change in proficiencies, Jordan's mean will be relatively smaller and France's mean relatively greater, after items are resolved. The UK mean might decrease slightly also.

To establish a common origin⁵ for comparing country mean proficiency estimates before and after resolving items, the mean difficulty estimates of the non-DIF items were

⁵ The estimates of person means before and after resolving items cannot be directly compared because each DIF analysis has its own origin as the item estimates in each of the two analyses sum to zero, despite the different total number of items in the two analyses. The procedure was carried out using a function in RUMM2030.

anchored to their values before any items were resolved and then the parameters of the resolved items were re-estimated. This ensured that both analyses had the same origin, permitting direct comparisons of mean proficiencies generated from the two analyses. The mean proficiencies and their standard deviations before and after resolving items for each country are displayed in Table 6.

<<INSERT TABLE 6 ABOUT HERE>>

Based on Table 6, the deviations of means before and after resolving items are negative for the UK and Jordan and positive for France. This finding is consistent with the expectations from the relative item difficulties of the resolved items indicated above. The differences in means and standard deviations, before and after resolving items, are small with only differences of person means of France and Jordan being significant at the 95% confidence interval. Mean proficiency estimates did not change materially because DIF balanced out at the test level. Hence, the tensions of trading model fit for invariance at test level and/or the validity of comparisons, seem to be solved empirically. Nevertheless, one could imagine a case with greater and statistically significant deviations in mean proficiencies before and after resolving items. Such deviations can have major implications on the ranking of countries and the validity of the comparisons. And even a small quantitative effect can be noticeable in rankings among countries (e.g., Kreiner & Christensen, 2014) on which policy makers seem focused (Baird et al., 2016; Steiner-Khamsi, 2003). Therefore, below we make concrete the trade-off between model fit and invariance of item difficulties as it affects the validity of comparisons.

The perspective of this trade-off is that in large-scale assessments, the variable of assessment is necessarily a combination of a causal and composite variable described earlier. To stress the point also made earlier, the source of DIF (which governs the decision to be made) can generally be identified only using information external to the analyses of the data in which the presence of DIF is found.

First, we illustrate a justification for *not* resolving items. For example, suppose there are items with real DIF that are relatively easier for a country and the items are resolved. Then the resolved items for that country will also have relatively easier difficulty estimates. As a consequence, a country will receive smaller rewards for correct responses to these items and the mean of the country will be reduced relative to its original estimate. However, if the items are relatively easier for the country because of better teaching and curriculum coverage of the aspects assessed by the items, then resolving the item does not reward the country for this effort. Accordingly, resolving is not justified. Thus, to enhance the validity of assessments and comparisons, not resolving the items and thereby compromising model fit for invariance, is justified.

Second, we illustrate a justification for resolving items. For example, suppose a similar situation as above but with a set of items with real DIF that are more difficult for a country. Then if the items are resolved, a correct response for students from that country will receive a relatively greater reward for a correct response. Now suppose that the source of DIF that makes the items more difficult for the country is irrelevant to the variable, such as poor translation of items in the language of that country. In this case, to overcome the effect of the poor translation for the country, the resolution of the items is justified. Simultaneously, those countries finding the items easier because they did not have a bad translation are not over-

rewarded. In this case, to enhance the validity of assessments and of comparisons, it is justified to compromise item invariance in order to improve model fit.

From the perspective of group comparisons, we note that deleting items has the same quantitative effect as resolving items and therefore whether to delete or retain an item involves the same principles as those of resolving an item – it depends on the source of DIF being irrelevant or relevant to the assessment of the variable. However, as noted earlier, from the perspective that the variable is in part a composite variable, the resolution of items ensures coverage of the curriculum. Then in addition to reducing reliability, deleting an item removes the assessment of the distinct aspect assessed by the item and thus reduces validity.

Summary

Common ways of handling DIF in items in assessments *post hoc* include deleting or resolving those items, or even ignoring the DIF. Each of these options could be valid depending on the source of DIF. One challenge, however, is that very often the source of DIF is unknown: there could be a myriad of different reasons that makes decisions about leaving items as they are, deleting them, or resolving them, hard to make. Each of these options has direct implications for model fit, invariance, and validity. We argued that very often the decision to adopt one option over the two others will be a trade-off between model fit and invariance of relative item difficulties, a condition typically required for making valid comparisons.

Deleting items exhibiting DIF improves model fit and can retain invariance of functioning of the remaining items, but by reducing the number of items, reliability is reduced. Further, by not assessing the distinct aspects of the variable assessed by the deleted items, the validity of the assessment of the variable is reduced. This feature is relevant to the

case of scientific literacy where the variable can be conceptualized as both a composite and a causal variable. Resolving items also improves model fit, but by retaining the items, maintains the validity and reliability of the assessments. However, because the resolved item results in group-specific difficulties for the item, the difficulty of the item is no longer invariant across countries. This method of accounting for DIF *post hoc* is justifiable when the source of DIF is irrelevant to the variable and creates unfair advantages or disadvantages in comparisons across countries. On the other hand, it is argued that if the source of DIF is relevant to the variable, then resolving items can be considered unjustifiable and maintaining invariance takes precedence over obtaining model fit.

The decision to resolve an item in making proficiency estimates, as distinct from using it as a means of identifying DIF, should therefore be driven by a careful content analysis of the item and an understanding of whether and how a particular construct is being taught in the advantaged and disadvantaged countries, or, in the case of items being rendered in different languages, the quality of the translations and idiosyncrasies of the languages involved. This information can only come from outside the data analyzed to obtain proficiency estimates.

In the illustrative study of this paper in which items with real DIF were resolved, the impact on proficiency estimates of countries was small. This effect was primarily because, even though the proportion of items showing real DIF was relatively large (22%), the variation in DIF among items was substantially compensated for within the countries. However, if DIF is not compensated for among items, then the decision to resolve items should have a substantive basis, even though identifying the source of DIF may be challenging (Andrich & Hagquist, 2015; Penfield & Camilli, 2007). If the source of DIF is the language used (e.g., El Masri et al., 2016), resolution would be a justifiable option and trading off invariance for model fit would be appropriate. If the source of DIF is different

curriculum coverage across the countries (e.g., Huang et al., 2016) and hence different emphasis on PISA scientific literacy competencies across the three countries, then ignoring DIF (i.e., neither deleting nor resolving DIF items) and prioritizing invariance over model fit would be the most appropriate option.

In this study, Jordan had a relatively lower overall proficiency mean in science compared to the other two countries despite finding the DIF items relatively easier. Because of limited access to the items (the content of only four out of the 23 DIF items were released to the public), it was not possible to carry out the necessary content analysis of the items to illustrate our argument. While recognizing the importance of protecting item security, we hope that researchers wishing to investigate this line of research would be granted access to the actual items of interest in the future.

References

- Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29, 67–91.
- Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42, 69–81.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–574.
- Andrich, D. (1988). *Rasch Models for Measurement*. Newbury Park, CA: Sage.
- Andrich, D. (2014). A structure of index and causal variables. *Rasch Measurement Transactions*, 28, 1475–1477.
- Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics*, 37, 387–416.
- Andrich, D., & Hagquist, C. (2015). Real and Artificial Differential Item Functioning in Polytomous Items. *Educational and Psychological Measurement*, 75, 185–207.
- Andrich, D., Sheridan, B. S., & Luo, G. (2015). RUMM2030: An MS Windows computer program for the analysis of data according to Rasch Unidimensional Models for Measurement. Perth, Western Australia: RUMM Laboratory.
- Asil, M., & Brown, G. T. L. (2016). Comparing OECD PISA Reading in English to other languages: Identifying potential sources of non- invariance. *International Journal of Testing*, 16, 71–93.
- Baird, J., Johnson, S., Hopfenbeck, T. N., Isaacs, T., Sprague, T., Stobart, G., & Yu, G. (2016). On the supranational spell of PISA in policy. *Educational Research*, 58, 121–138.

- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, *Statistical theories of mental test scores* (pp. 397–545). Reading, Mass.: Addison-Wesley.
- Duncan, O. D. (1984). Rasch measurement: Further examples and discussion. In C. F. Turner & E. Martin (Eds.), *Surveying Subjective Phenomena* (Vol. 2, pp. 367–401). New York: Russel Sage Foundation.
- Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56, 52–64.
- Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, 5, 155–174.
- El Masri, Y. H., Baird, J., & Graesser, A. C. (2016). Language effects in international testing: The case of PISA 2006 science items. *Assessment in Education: Principles, Policy & Practice*, 23, 427–455.
- Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179–197.
- Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, 5, 23–35.
- Fischer, G. H. & Molenaar I. W. (1995). *Rasch models: Foundations, recent developments, and applications* (pp. 3–14). New York, NY: Springer.
- Grisay, A., de Jong, J. H. A. L., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement*, 8, 249–266.
- Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. *IERI Monograph Series:*

Issues and Methodologies in Large-Scale Assessments, 4, 63–83.

Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation*, 33, 69–86. Hagquist, C., & Andrich, D. (2017), Recent Advances in Analysis of Differential Item Functioning in Health Research using the Rasch Model. *Health and Quality of Life Outcomes*, 15, 181 – 188.

Hauger, J. B., & Sireci, S. G. (2008). Detecting differential item functioning across examinees tested in their dominant language and examinees tested in a second language. *International Journal of Testing*, 8, 237–250.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Huang, X., Wilson, M., & Wang, L. (2016). Exploring plausible causes of differential item functioning in the PISA science assessment: language, curriculum or culture. *Educational Psychology*, 36, 378–390.

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika*, 79, 210–231.

Le, L. T. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing*, 9, 122–133.

Millsap, R. E., & Everson, H. T. (1993). Methodology review : Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, 17, 297–333.

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, 87, 224–240.

OECD. (1999). *Measuring Student Knowledge and Skills: A New Framework for Assessment*.

Paris: OECD Publishing.

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

2006. Paris: OECD Publishing.

OECD. (2009). *PISA 2006 Technical Report*. Paris, France: OECD.

OECD. (2017). PISA 2015 science framework. In *PISA 2015 Assessment and Analytical*

Framework (pp. 19–49). Paris: OECD Publishing.

Oliveri, M. E., Olson, B. F., Ercikan, K., & Zumbo, B. D. (2012). Methodologies for

investigating item- and test-level measurement equivalence in international large-scale assessments. *International Journal of Testing*, 12, 203–223.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R.

Rao & S. Sinharary (Eds.), *Handbook of Statistics: Psychometrics* (Vol. 26, pp. 125–167). North Holland: Elsevier.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment*

Tests (Copenhagen, Danish Institute for Educational Research.) (Vol. Expanded e).

Chicago: The University of Chicago Press.

Sandilands, D., Oliveri, M. E., Zumbo, B. D., & Ercikan, K. (2013). Investigating Sources of

Differential Item Functioning in International Large-Scale Assessments Using a Confirmatory Approach. *International Journal of Testing*, 13, 152–174.

Steiner-Khamsi, G. (2003). The politics of league tables. *Journal of Social Science*

Education, 1, 1–6.

Stenner, J., Burdick, D. S., & Stone, M. H. (2008). Formative and reflective models: can a

Rasch analysis tell the difference? *Rasch Measurement Transactions*, 22, 1152–1153.

- Tesio, L. (2014). Items and variables, thinner and thicker variables: Gradients, not dichotomies. *Rasch Measurement Transactions*, 28, 1477–1479.
- Van der Linden, W. J. (Ed.). (2016). *Handbook of item response theory* (Vol. 1). Taylor and Francis. Boca Raton, Florida. USA.
- Wang, W., & Su, Y. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Psychological Measurement*, 17, 113–114.
- Wang, X., Bradlow, T., Wainer, H., & Muller, E. S. (2008). A Bayesian method for studying DIF: A cautionary tale filled with surprises and delights. *Journal of Educational and Behavioral Statistics*, 33, 363–384.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427–450.
- Wu, A. D., & Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing*, 6, 287–300.
- Xie, Y., & Wilson, M. (2008). Investigating DIF and extensions using an LLTM approach and also an individual differences approach: An international testing context. *Psychology Science Quarterly*, 50, 403–416.
- Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4(2), 223–233.

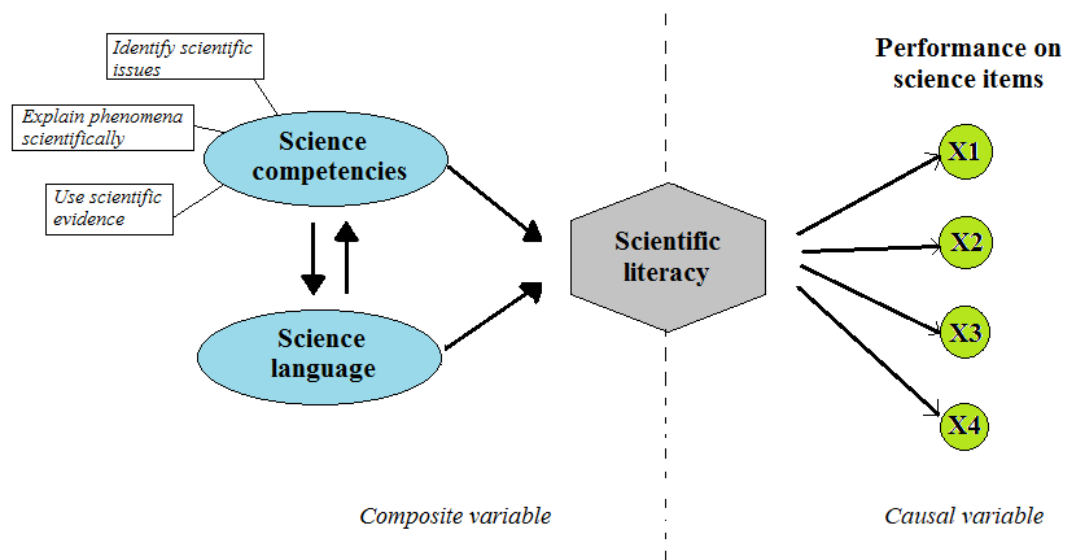


Figure 1. Causal and composite nature of scientific literacy in PISA.

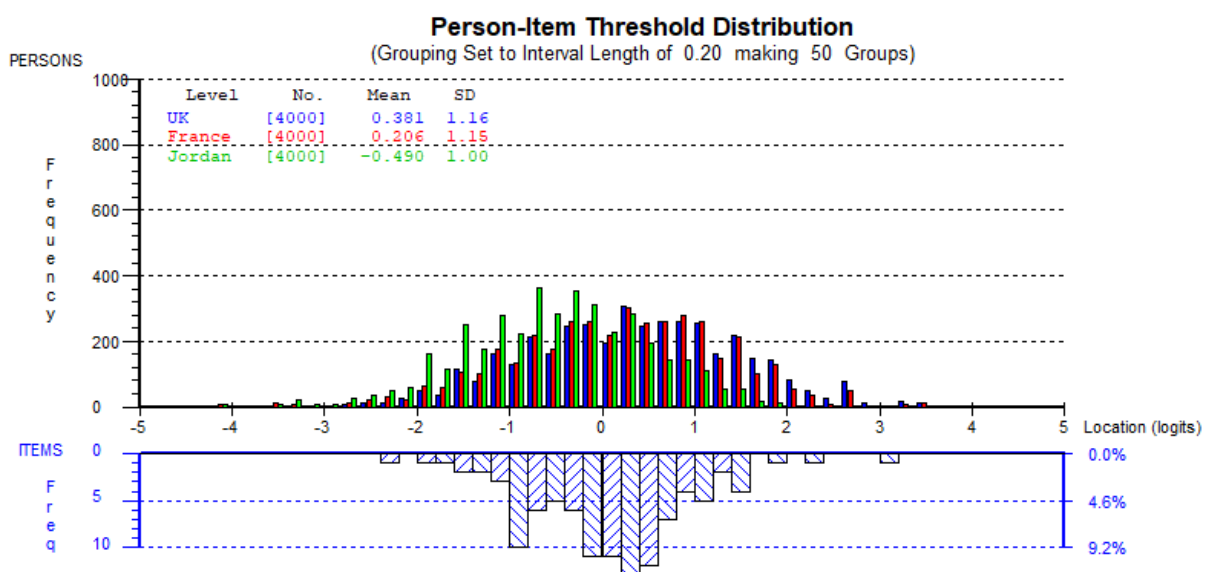
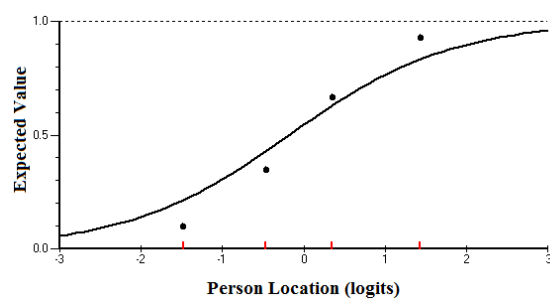
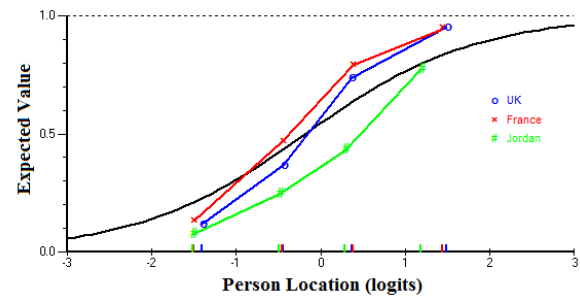


Figure 2. Distribution of persons and items across three countries in PISA 2006 science.

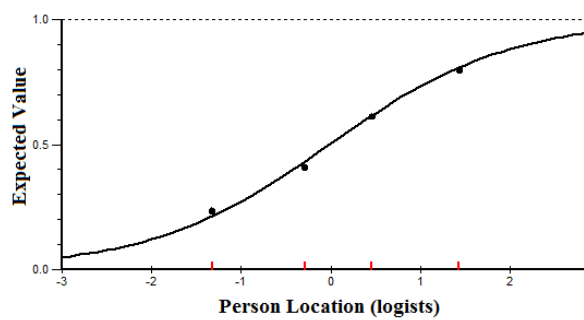


Overall sample

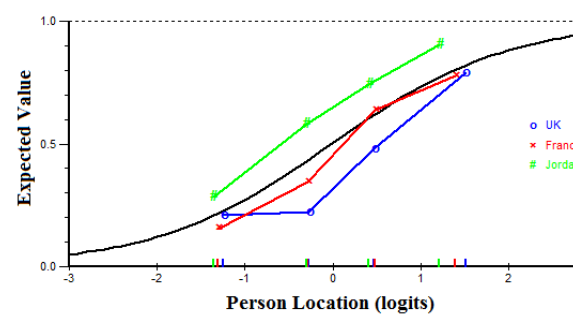


By country

Figure 3a. Means in class intervals relative to the ICC of Item 1.



Overall sample



By country

Figure 3b. Means in class intervals relative to the ICC of Item 43.

Table 1

Number of participants per country in the initial and reduced samples

	UK	France	Jordan	Total
Initial Sample	13 152	4 716	6 509	24 377
Analysis Sample	4 000	4 000	4 000	12 000

Table 2

Two-way ANOVA of residuals of four class intervals

Item	PISA code	<u>Class Interval</u>			<u>Uniform DIF</u>		
		MS	F	P	MS	F	P
1	S114Q03T	61.890	12.912	0.000000	87.632	18.283	0.000000
9	S304Q03A	22.219	3.810	0.101570	88.645	15.202	0.000005
15	S416Q01	6.368	0.974	0.404330	85.227	13.046	0.000013
18	S425Q03	46.326	5.888	0.000588	283.22	36.002	0.000003
23	S458Q01	17.96347	2.640	0.048888	253.54	37.264	0.000008
30	S519Q03	37.67606	4.774	0.002741	88.161	11.171	0.000016
41	S256Q01	45.03914	5.016	0.001964	97.342	10.840	0.000020
43	S268Q06	0.85333	0.120	0.948239	59.560	8.385	0.000255
50	S408Q04	1.66938	0.228	0.876769	85.512	11.689	0.000008
51	S413Q04T	4.50771	0.617	0.604078	61.208	8.381	0.000262
54	S415007T	7.22752	1.109	0.345207	86.408	13.253	0.000009
58	S426Q03	12.18427	1.749	0.156919	209.15	29.952	0.000006
59	S426Q05	19.81863	3.438	0.016816	83.033	14.405	0.000001
62	S428Q03	7.93073	1.256	0.288788	54.525	8.637	0.000217

Table 2 (continue)

Two-way ANOVA of residuals of four class intervals

Item	PISA code	<u>Class Interval</u>			<u>Uniform DIF</u>		
		MS	F	P	MS	F	P
63	S437Q01	11.078	1.515	0.209743	78.111	10.683	0.000018
67	S438Q02	14.868	1.918	0.125777	143.56	18.518	0.000001
76	S466Q07T	2.524	0.361	0.781051	120.34	17.224	0.000000
86	S493Q01T	4.383	0.596	0.618140	72.204	9.812	0.000054
90	S495Q04T	1.386	0.196	0.898821	69.525	9.856	0.000066
94	S508Q03	9.344	1.574	0.194884	75.943	12.789	0.000005
95	S510Q01T	39.588	4.819	0.002572	96.578	11.757	0.000008
96	S519Q02T	108.790	12.169	0.000009	88.084	9.853	0.000064
103	S485Q02	7.759	1.180	0.316745	53.79768	8.18277	0.000309

MS: mean square, F: F-statistic; P: Probability

Table 3

Likelihood values of dataset before and after resolving DIF items

Analysis	Log Likelihood	Parameters estimated
Initial	-201368.12	109
Resolving DIF items	-198789.58	155
Chi Square	5157.07	46

Table 4

Item difficulty of DIF items and average difficulty per country after resolving

Item	PISA code	<u>Item location</u>			
		<u>Before resolving</u>	<u>After resolving</u>		
		Initial sample	UK	France	Jordan
1	S114Q03T	-0.178	-0.430	-0.620	0.530
9	S304Q03A	0.447	1.260	0.230	-0.150
15	S416Q01	0.879	0.590	0.670	1.800
18	S425Q03	0.328	1.170	0.670	-0.840
23	S458Q01	1.418	1.840	2.810	0.070
30	S519Q03	1.443	2.240	1.430	0.800
41	S256Q01	-1.960	-2.900	-1.150	-2.280
43	S268Q06	-0.013	0.420	0.220	-0.550
50	S408Q04	0.738	1.530	0.430	0.400
51	S413Q04T	0.668	0.120	1.130	0.840
54	S415007T	-0.896	-1.150	-1.340	-0.260
58	S426Q03	-0.897	-1.530	0.070	-1.690
59	S426Q05	-0.860	-0.960	-1.470	-0.260
62	S428Q03	-0.411	-1.170	0.250	-0.390

Table 4 (continued)

Item difficulty of DIF items and average difficulty per country after resolving

Item	PISA code	<u>Item location</u>			
		<u>Before resolving</u>	<u>After resolving</u>		
		Initial sample	UK	France	Jordan
63	S437Q01	-0.921	-0.460	-0.500	-1.700
67	S438Q02	-0.577	-1.170	0.130	-0.850
76	S466Q07T	-0.700	-1.060	-1.250	0.100
86	S493Q01T	0.520	-0.040	0.840	0.850
90	S495Q04T	0.305	0.100	-0.110	1.090
94	S508Q03	-1.040	-1.320	-0.220	-1.600
95	S510Q01T	0.055	-0.100	0.780	-0.370
96	S519Q02T	0.230	-0.130	0.870	0.000
103	S485Q02	-0.235	-0.340	0.410	-0.710
	Mean	-1.657	-0.152	0.186	-0.225

Table 5

Within-subject effects (repeated-measures ANOVA) at 5% significance level

	Sum of Squares	df	Mean Square	F	Significance
Item location	2.210	2	1.105	2.137	0.130
Error	22.753	44	0.517		

df: degrees of freedom; F: F-statistic

Table 6

Proficiency means before and after resolving items with fixed origin together with estimated confidence intervals

		<u>Before resolving</u>			<u>After resolving¹</u>			Difference in means
	Sample	Mean	SD	CI	Mean	SD	CI	
UK	4000	0.38	1.16	0.36 – 0.40	0.36	1.2	0.34 – 0.38	-0.02
FRA	4000	0.21	1.15	0.19 – 0.23	0.26	1.16	0.24 – 0.28	0.05*
JOR	4000	-0.49	1.00	-0.51 – (-0.47)	-0.53	1.02	-0.55 – (-0.51)	-0.04*

¹*Estimates based on the analysis where the origin was fixed; SD: standard deviation; CI: confidence interval*

* *statistically significant at 95% confidence interval*

