Contributions to latent variable modeling in educational measurement

Zwitser, R.J.

Chapter 4

Monitoring Countries in a Changing World. A New Look at DIF in International Surveys

Summary

This paper discusses the issue of differential item functioning (DIF) in international surveys. DIF is likely to occur in international surveys. What is needed is a statistical approach that takes DIF into account, while at the same time allowing for meaningful comparisons between countries. Some existing approaches are discussed and an alternative is provided. The core of this alternative approach is to define the construct as a large set of items, and to report in terms of summary statistics. Since the data are incomplete, measurement models are used to complete the incomplete data. For that purpose different models can be used across countries. The method is illustrated with PISA's reading literacy data. The results indicate that this approach fits the data better than the current PISA methodology, however, the league tables are nearly the same. The implications for monitoring changes over time are discussed.

This chapter has been submitted for publication as: Zwitser, R.J., Glaser, S. & Maris, G. (submitted). Monitoring Countries in a Changing World. A New Look at DIF in International Surveys.

4.1 Introduction

Since a couple of decades, educational surveys have been administered repeatedly with the purpose to explore what students know or can do, to compare participating countries or economies, to measure trends over time, and/or to evaluate educational systems. Examples of such surveys are the *Programme for International Student Assessment* (PISA), the *Trends in International Mathematics and Science Study* (TIMSS), the *Progress in International Reading Literacy Study* (PIRLS), and the *European Survey on Language Competences* (ESLC).

The constructs that are supposed to be measured in these surveys are operationalized in sets of items. These sets are usually too large to administer completely to each student in the sample. Therefore, the survey is administered in an *incomplete design*, in which only subsets of items are administered to each student. This implies that there are structural missing data. In order to get comparable scores from incomplete data, surveys make use of latent variable models (see e.g., Lord & Novick, 1968, for a general introduction). The models, of which the parameters can be estimated with the observed incomplete data, describe the distribution on the total set of items.

To obtain unbiased results, it is important to find a model that fits the data. However, an issue that complicates the statistical modeling, is the occurrence of *differential item functioning* (DIF). Holland & Wainer (1993) define DIF as follows: 'DIF is a relative term. An item may perform differently for one group of examinees relative to the way it performs for another group of examinees'. In this paper, we focus on two types of DIF. The first is uniform DIF, which, in terms of latent variable models, means that conditional on the latent variable, the probability of a particular item response varies across subpopulations. The second is non-uniform DIF, which means that the correlation between a particular item response and the latent variable varies across subpopulations.

It has been demonstrated more than once that there is DIF in educational surveys (Kreiner, 2011; Kreiner & Christensen, 2013; Oliveri & Ercikan, 2011; Oliveri & Von Davier, 2011; Oliveri & Von Davier, 2014; OECD, 2009a). The presence of DIF is usually seen as a threat to validity, and as something that limits score comparability between subpopulations (American Educational

Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). In this paper, however, we want to emphasize that for surveys DIF could be one of the main interesting outcomes, and that it need not invalidate meaningful comparisons between countries. The purpose of this paper is to describe and illustrate such a method.

The paper is structured as follows. First, we discuss some of the existing approaches concerning DIF and surveys (Sections 4.1.1, 4.1.2, and 4.1.3). Then, in Section 4.2, the new method is proposed and compared to the existing ones. Throughout the paper, we take the reading literacy items of PISA as an example. In Section 4.3 we describe two example data sets, whereafter in Section 4.4 three examples of our methodology are provided. The results and implications are finally discussed in Section 4.5.

4.1.1 Remove DIF items and ignore DIF in the model

The current practice in PISA is a two-stage calibration procedure (OECD, 2009a, chapter 9). First, a calibration is performed in each country[1]. Based on this calibration, items with poor psychometric properties (the so-called *dodgy items*) are marked. One of the criteria to mark an item as dodgy is if the item difficulty in a particular country is significantly lower or higher than the average of all available countries. If an item is dodgy in more than 10 countries, then it is removed from the survey, otherwise it is scored as 'not administered' in the countries in which it performs dodgy (OECD, 2009a, chapter 9). Therefore, it could be said that PISA attempts to remove DIF with respect to item difficulty. Then, in the second stage, PISA applies the Mixed Coefficients Multinomial Logit Model (Adams, Wilson, & Wang, 1997; OECD, 2009a) assuming a common scale for all countries together. This model is a multidimensional generalization of the Partial Credit Model (PCM), which is a Rasch model for partial credit scoring (Masters, 1982). The person parameters, which are in this case a monotone transformation of the expected sum score on the total set of items, are the basis for comparisons between countries. The use of a common scale for all countries together

[1]In fact, PISA consists of participating economies. However, since most economies are countries, and since we think that the term countries is easier for the reader, we use the term countries instead of economies.

implies that DIF is currently not taken into account in the psychometric model. It has been demonstrated that the procedure described above does not succeed in removing all DIF in PISA (Kreiner, 2011; Kreiner & Christensen, 2013; Oliveri & Von Davier, 2011; Oliveri & Von Davier, 2014; Oliveri & Ercikan, 2011). In order to improve the model fit and to study the effect of DIF on PISA's final scores and rankings, different alternative methods have been proposed. These alternatives are discussed in the following two sections.

4.1.2 Add subpopulation-specific item parameters and compare person parameter estimates

To study the score scale comparability in international assessments, Oliveri & Von Davier (2011; 2014) adjusted the measurement model with additional, country-specific item parameters. In both papers, they showed that the additional item parameters had a substantial positive effect on the model fit, however, the influence of the model adjustments on the final scores was limited. For all three domains (i.e., Science, Reading, and Mathematics) the correlation between the country means based on the international parameters (i.e., the PISA approach) and the country means based on partially-unique country parameters is at least 0.987 (Oliveri & Von Davier, 2011; 2014).

A problem with this approach, however, is that the (equated) person parameters are used as final scores, while for these scores the relationship (e.g., the ordering) between countries depends on an arbitrary decision about the scaling method. This is illustrated in the following example.

Consider the PCM in the following parameterization:

$$P(X_i = j|\theta) = \frac{\exp(j\theta - \sum_{g=1}^{j} b_{ig})}{1 + \sum_{h=1}^{m_i} \exp(h\theta - \sum_{g=1}^{h} b_{ih})} \quad (j = 0, \cdots, m_i), \tag{4.1}$$

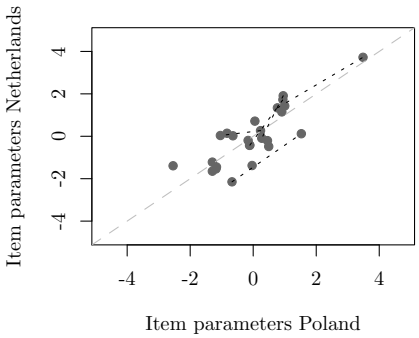
in which it is assumed that the response to item i, denoted by X_i, falls in the score range $\{0, 1, \cdots, m_i\}$, and where $b_{ig}, g = 1, \cdots, m_i$, are the parameters of item i. The example consists of 20 reading literary items (15 dichotomous, and 5 partial credit with possible scores $\{0, 1, 2\}$) of PISA 2006 and two particular countries: Poland and The Netherlands. Further details about the data are described in Section 4.3.2.

Let us assume that item covariance structures are different in Poland and

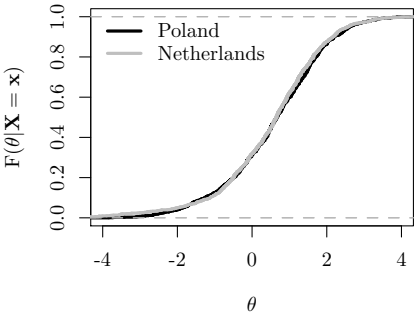
The Netherlands. In that case, country specific parameters are needed in the model. In this example, we choose to estimate separate PCMs on the data of Poland and The Netherlands, respectively. What we obtain are two sets of parameters that are not on the same scale. How to equate these scales? Many answers to this question have been considered (see, for instance, Kolen & Brennan, 2004, chapter 6). However, to illustrate the in principle arbitrary aspect of scaling with DIF, we just consider the following three options.

The first option is to fix the means of both sets of item parameters at the same value, e.g., at zero. This *mean-mean* method (Loyd & Hoover, 1980) was also applied by Oliveri & Von Davier (2011; 2014). The corresponding scatterplot with item parameters[2] is depicted in Figure 4.1a. Observe that the dots are not on an approximately straight line, which indicates that there is DIF between these two countries. The cumulative distribution of person parameters is displayed in Figure 4.1b. The θ-distributions are approximately equal.

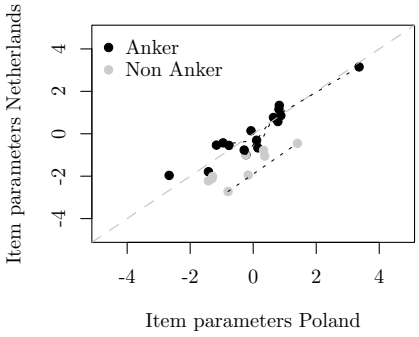
[2]The parameters of polytomous items are connected with a dotted line.



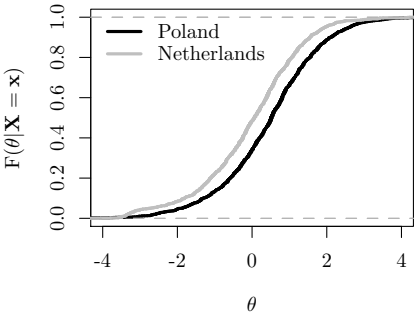
(a) Item parameters - option 1



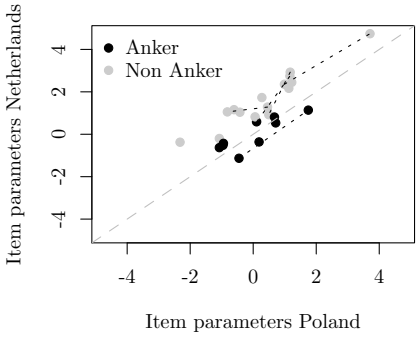
(b) Person parameters - option 1



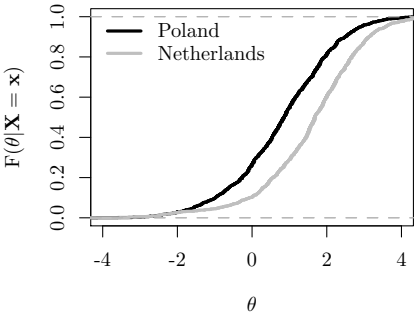
(c) Item parameters - option 2



(d) Person parameters - option 2



(e) Item parameters - option 3



(f) Person parameters - option 3

Figure 4.1: Different anchoring options.

The second and third option are based on equating with an anchor set of items. In the presence of DIF, the choice of an anchor is problematic. There can be clusters of items, such that within a cluster the relative difficulties are invariant across populations (Bechger, Maris, & Verstralen, 2010; Bechger & Maris, 2014). An example of such clusters is shown in Figures 4.1c en 4.1e. The items depicted with black and grey circles, respectively, could be seen as two of such clusters. It is not uncommon in psychometrics to take one cluster as anchor and consider these items as the items *without DIF*. The remaining items are then considered to be the *DIF items*. In this example we have two clusters, but taking one or the other as anchor (denoted as options 2 and 3), or fixing the mean of the item parameters (option 1) is statistically equivalent. These three options have exactly the same likelihood, and are only a re-parameterization of each other. This implies that there is no empirical evidence to prefer one option over the other. Observe, however, that the relative difference between person parameter distributions is different across these three options. The θ-distributions related to options 2 and 3 are depicted in Figures 4.1d and 4.1f, respectively. With option 2 the distribution of Poland is stochastically *larger* than the distribution of The Netherlands, while the reverse is true with option 3. Since the ordering of the person parameter distribution of person parameters depends on an arbitrary decision about the scaling method, we argue that person parameters are in this case not a suitable basis for comparing countries.

4.1.3 Add subpopulation-specific item parameters and adjust the observed total score

Kreiner & Christensen (2007) also came up with a model with subpopulation-specific item parameters. Since they also doubted whether comparisons of person parameters in such models are meaningful, they suggested to estimate adjusted true scores on a specific set of items. Their procedure implies that for each focus group the estimated true score is adjusted as if the members of the focus group were members of the reference group. This procedure was applied in their paper about the reading items of PISA 2006 (Kreiner & Christensen, 2013). It was observed that the average number of correct items in Belgium was 15.09, but if these Belgian students would have been from Azerbaijan, they would have scored 17.59 (see Kreiner & Christensen, 2013, Table A.1.). Is it meaningful to adjust observed scores based on a DIF analysis? We think

that the answer depends on the type of DIF. If there is an external source that has a temporary effect on the scores, and the magnitude of the effect can empirically be estimated, then it makes sense to adjust test scores. For instance, wind assistance in a 100 meter sprint exists independent of the ability of the athlete, and benefit of wind assistance can be modeled explicitly (see e.g., Linthorne, 2014). Since it could have happened that the same athlete would have ran the race without wind, the adjusted sprint time is a meaningful measure. In a testing context, however, the sources of DIF are in most cases unclear (Sandilands, Oliveri, Zumbo, & Ercikan, 2013; American Educational Research Association et al., 1999). Sources that are reported in a survey context (e.g., language, cultural, or gender differences) are in a fixed way related to the subpopulation. They are fixed in the sense that it could not have happened that the same student was a girl instead of a boy, or that he was English speaking instead of French speaking. Therefore, we think that DIF-based score adjustments like 'if this student was a member of another subpopulation' have a limited meaning in a survey context.

4.1.4 DIF as an interesting outcome

Besides the papers mentioned in Section 4.1.1, PISA itself also reports examples of DIF. For instance, with respect to the math items of PISA 2003, *"among OECD countries, the Slovak Republic ranks around fourteenth (twelfth to seventeenth) and thirteenth (ninth to seventeenth) for the 'space and shape' and 'quantity' scales, but around twenty-fourth (twenty-fourth to twenty-fifth) in the 'uncertainty' scale"* (OECD, 2004). Since opposite patterns occur in other countries (OECD, 2004), one can easily conclude that there is DIF between countries with respect to the math items. We believe that this kind of DIF should not be eliminated. Instead, it should be considered as an interesting outcome of an educational survey. Therefore, the incomplete survey data should be treated in a way that takes DIF into account. The previous sections, however, illustrate that the existing approaches in which DIF is taken into account are not fully satisfactory. In the next section, we propose another DIF driven approach to model the data obtained from international surveys.

4.2 Method

The main idea is straightforward: define the construct as a large set of items, collect the data in an incomplete design, use measurement models to complete the incomplete data, and report in terms of summary statistics. The following sections describe this procedure in more detail.

4.2.1 The construct

The starting-point of our approach is that the subject domain of interest is defined by a large set of items. This is in accordance with the third variation of the *market basket* approach (Mislevy, 1998). What follows is that the primary interest is in knowing how students perform on this set of items. Statistically speaking, we want to estimate the distribution of $\mathbf{X}$, which represents the matrix with responses from all students to all items. What we therefore need to have, is an unbiased estimate $\mathbf{x}$. One way to achieve this, is to administer the complete set of items to a simple random sample of students. However, this is practically untenable. The current practice is to draw a stratified sample and to administer the survey with an incomplete design. The correction for stratified sampling is already taken into account with the student weights and replicate weights (OECD, 2009b, chapter 3 and 4). What remains is how the incomplete observed data may be used to get an unbiased estimate of the complete data. This is the point where the measurement model comes in.

4.2.2 Purpose of the measurement model

The primary role of the model is to describe the scores on the missing data, conditionally on the observed data. If the model fits, then the estimated complete matrix $\mathbf{x}$ is an unbiased estimate of $\mathbf{X}$. Observe that for this purpose it is not required that the same model is used for each country. Instead, it is possible to have different models for different countries, or even to have different models for different subpopulations within a country. It only needs to be verified whether the model(s) is (are) suitable to impute missing data.

4.2.3 Comparability

One of the main applications of surveys is to compare performance across participating countries. This implies the comparison of subsets of $\mathbf{X}$. Let us denote the subsets corresponding to countries a and b as $\mathbf{X}_a$ and $\mathbf{X}_b$. Under the conditions described above, $\mathbf{x}_a$ and $\mathbf{x}_b$ are unbiased estimates of $\mathbf{X}_a$ and $\mathbf{X}_b$, respectively, and are therefore comparable between countries. Consequently, the same holds for functions applied to $\mathbf{X}$. If one is interested in the function f of $\mathbf{X}$ (e.g., the sum score over items), then $f(\mathbf{x}_a)$ and $f(\mathbf{x}_b)$ are unbiased estimates of $f(\mathbf{X}_a)$ and $f(\mathbf{X}_b)$. Notice that this only holds if the same function f is applied to both $\mathbf{x}_a$ and $\mathbf{x}_b$. With two *different* functions, e.g., $f_1(\mathbf{x}_a)$ and $f_2(\mathbf{x}_b)$, the comparability property is lost.

4.2.4 Difference with existing methods

It could be that one of the models to which is referred in Sections 4.1.2 and 4.1.3 does fit the data. In that case, these models fit in our approach because they can be used to estimate the complete data, and consequently also summary statistics on the complete data. The difference between our approach and the methods described in Sections 4.1.2 and 4.1.3 is about what is done with the estimated complete data. Oliveri & Von Davier (2011; 2014) use the model to compare the estimated person parameters. However, the estimated person parameter is a function of the data, and the item parameters. If the item parameters differ in different subpopulations, the comparison of person parameters is based on two different functions of the data (cf., f_1 and f_2). As we have explained in the section above, scores based on different functions of the data are not comparable.

Kreiner & Christensen (2013) illustrate their method with complete data, and use the model to adjust the total scores. This implies that they compare the observed performance of country a with the adjusted performance of country b. In contrast, the method that we propose does not contain such adjustments. If we have complete data, we compare the performance of country a with the (non-adjusted) performance of country b.

4.2.5 Estimation process

For the method described above, it is needed to find a subdivision of the population, such that for each subpopulation a suitable model can be used. To simplify the illustrations, this paper only verifies the application of different models for different countries and for different time points in the survey. To stay close to illustrations provided in the related research that was discussed in Section 4.1, two models are taken under consideration: the PCM as in (4.1), and the *One-Parameter Logistic Model* (OPLM, Verhelst & Glas, 1995):

$$P(X_i = j|\theta) = \frac{\exp[a_i(j\theta - \sum_{g=1}^{j} b_{ig})]}{1 + \sum_{h=1}^{m_i} \exp[a_i(h\theta - \sum_{g=1}^{h} b_{ih})]} \quad (j = 0, \cdots, m_i),$$

which compared to (4.1) also contains integer a_i-parameters that are considered as being known before. These are also called *discrimination indices*.

For both the PCM and the OPLM, the b-parameters are estimated with the conditional likelihood method (Andersen, 1973a). An important question related to the OPLM is how to obtain these a_i-parameters. Observe that items with the same value for the a_i-parameter do form a Rasch subscale. Therefore, an OPLM can also be considered as a collection of Rasch subscales. One way to estimate this model, is to 1) find the different Rasch subscales, and 2) investigate how these are related to each other, e.g., find the different a_i-values (see Bolsinova, Maris, and Hoijtink (2012) for recent developments with respect to this approach). For this paper, the a_i-parameters are estimated with the OPCAT procedure in the OPLM program (Verhelst et al., 1993). Since this is also an estimation procedure, the sample is split randomly into two subsamples. The first subsample, which is chosen to be approximately twice as large as the second, is used for estimating the a_i-parameters. The second subsample is used for estimating the b_i-parameters of the OPLM, while considering the a_i-parameters as known, and for evaluating the model fit. This estimation is performed with the OPLM program (Verhelst et al., 1993).

When the item parameters are estimated, then the next step is to estimate person parameters. In order to do so, five plausible values (PV) are drawn from the posterior distribution of the person parameters, conditional on the data and the estimated item parameters. The sample from this posterior distribution

can be drawn with a Gibbs sampler, in which the mean and the variance of the prior normal distribution are also estimated. A general description of this method can be obtained from the ESLC Technical Report (Council of Europe, 2012). Detailed discussions about the estimation algorithm (Marsman, Maris, Bechger, & Glas, 2013a), and about the advantages of using PV (Marsman, Maris, Bechger, & Glas, 2013b) are elsewhere available.

4.2.6 Plausible responses and plausible scores

As mentioned before, the person parameters are not directly comparable between countries if the sets of item parameters are not the same across these countries. However, the PV can be used to obtain *plausible responses*. Plausible responses are samples from the item response distribution according to the measurement model, the estimated item parameters for the particular country, and the PV for the person parameters in that country. For each PV, one set of plausible responses was drawn for the parts of the data that were missing. A full matrix with observed and plausible responses for each student on each item is an estimate of $\mathbf{X}$. It is already noticed in Section 4.2.3 that someone might be interested in statistics computed on $\mathbf{X}$. In this paper, we consider the sum over item responses (observed and plausible), and call these *plausible scores*. However, also other summary statistics could be taken. In Section 4.5, we will discuss that in this approach there are neither right nor wrong summary statistics.

4.2.7 Model fit evaluation

This approach relies upon a fitting model. In this paper, the fit is evaluated with two methods. The first method is the R_{1c} statistic. This statistic is, under the null-hypothesis that the model fits, asymptotically chi-square distributed (Verhelst et al., 1993). The second method is the exploration of item fit plots. These plots provide the theoretical ($\pi_{i|x_+}$) and observed ($P_{i|x_+}$) probability of a particular response (e.g., the correct response) on item i, conditional on the observed (weighted) sum scores (i.e., the sufficient statistic for the person parameter, from here on denoted as x_+). The observed x_+ are first binned into score groups, such that each score group had a substantial number of observations. Then the weighted average value of $\pi_{i|x_+}$ and $P_{i|x_+}$ are computed

Table 4.1: Sample sizes data set 1.

Year	Canada	Mexico
2003	1097	592
2006	1738	2085

in each score group, using the observed frequency of scores in the score group.

4.3 Data

In order to illustrate the proposed method, PISA's 'reading literacy' scale, is taken as an example[3]. For this construct, the PISA cycles of 2003 and 2006 contain the same 28 items. Although PISA does not claim that this set of items *defines* reading literacy, we do so to illustrate the method.

4.3.1 Data set 1

For the first two examples, only the data of booklet 11 of PISA 2003 and booklet 6 of PISA 2006 were taken, because these booklets contain all 28 reading literacy items, and can therefore be considered as complete data. Within these booklets, two particular countries with a large sample sizes are selected: Canada, and Mexico. Within these countries, only the cases without incidental missing values are selected. The sample sizes of the resulting data set are displayed in Table 4.1.

4.3.2 Data set 2

For the third example all reading literacy data of PISA 2006 were taken. The total sample consists of 398,750 students. For the example, the following data cleaning was performed. First, all students with more than 50% missing item responses within the administered booklet were removed from the data (i.e., 4,432 students $\approx$ 1%). Then, two countries were excluded: the USA, because no reading items were administered, and Liechtenstein, because of the very small sample size. The sample sizes per country are displayed in the third and fourth column of Table 4.4. The third column denotes the remaining sample

[3]The data were retrieved from `http://pisa2003.acer.edu.au/downloads.php` and `http://pisa2006.acer.edu.au/downloads.php` on August 22nd, 2013.

after the exclusions described above. The fourth column denotes the number of students that took reading items.

Since the construct is defined in a set of items, and not all reading items are administered in each country (OECD, 2009a, Tabel 12.5), only the subset of 20 items was taken that was administered in each country. These are the same items as those that were taken by Kreiner and Christensen (see Kreiner & Christensen, 2013, Table 3).

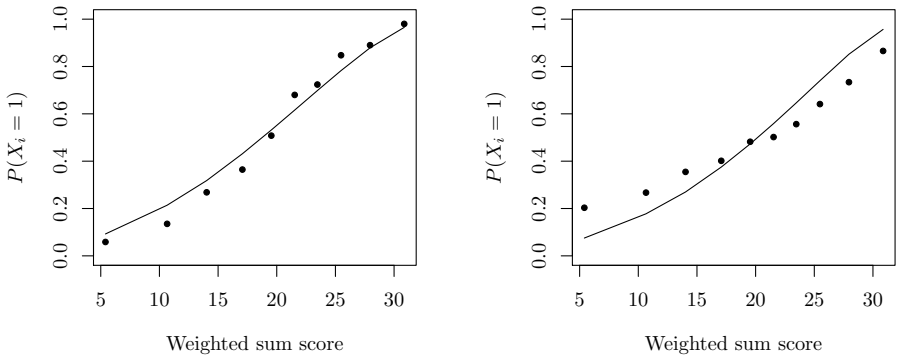
Treatment of missing data

PISA distinguishes three types of missing data: *not administered*, which not only means that items are not administered to a person, but also could mean that an items is removed due to *poor psychometric qualities*; *invalid response*, which implies a missing or double response; and *not reached*, which denotes consecutive invalid responses at the end of the test. Invalid and not reached response codes are considered as wrong (i.e., scored as 0). The coding of the not administered items needs some more explanation.

All reading items are distributed among 14 booklets. In some countries, some items are not administered in a particular booklet (see OECD, 2009a, Table 12.5). These cases are not considered as wrong responses, but literally as not administered. This implies that in the OPLM program (Verhelst et al., 1993) additional booklets are defined. These booklets are equivalent to the corresponding original booklet, besides the not administered item. After doing that, the data still contain not administered responses. For these cases, the following rule was defined: for countries where a particular item has more then 5% not administered responses, an additional booklet was defined for those students who have a not administered response on that particular item. This was the case for item R220Q06 in Estonia (multiple booklets) and item R227Q02T in Azerbaijan (only booklet 13). In cases where a particular item has less then 5% not administered responses, the responses were considered as wrong, i.e., scored as 0.

4.4 Illustrations and results

This section describes the method with three of illustrations. The first two examples are based on data set 1. In the first example (Section 4.4.1) the fit of



(a) Item 17 (b) Item 25

Figure 4.2: $P_{i|x_+}$ (dots), and $\pi_{i|x_+}$ (line) under the PCM for both countries and both cycles.

different measurement models is explored. The second example (Section 4.4.2) illustrates the observed sum score distribution of Canada and Mexico, as well as the estimated sum score distribution based on incomplete data. The last example (Section 4.4.3) is based on data set 2, and illustrates the impact of DIF on the average sum scores of countries, and the corresponding rankings.

4.4.1 Exploring the model fit

The same model for both countries

First, the PCM is fitted on the total data set with both countries and both cycles. This is actually what PISA also does. It turns out that this PCM does not fit, $R_{1c}(231) = 2{,}233.36$, $p < 0.0001$.

The item fit plots in Figure 4.2 provide a closer look at the fit of the PCM. The two examples of item 17 and 25[4] indicate that the addition of discrimination parameters would improve the model fit significantly. Therefore, the OPLM (Verhelst & Glas, 1995) is fitted on the total data set with both countries and both cycles. For this analysis, the data for each

[4]The item numbering is according to the order in which the items appear in booklet 6 of PISA 2006.

Table 4.2: Size subsamples and fit statistics OPLM.

Year	Country	Sample size		Fit statistics	
		subsample 1	subsample 2	$R_{1c}(99)$	p
2003	Canada	700	397	105.932	0.2985
	Mexico	400	192	122.306	0.0561
2006	Canada	1200	538	131.750	0.0155
	Mexico	1400	685	150.115	0.0007

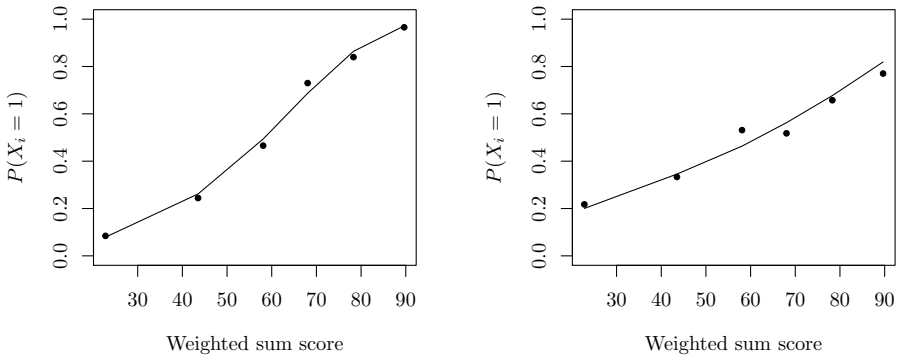
Table 4.3: Estimated a-parameters of the OPLM for both countries and both cycles.

Item nr.	a	Item nr.	a	Item nr.	a	Item nr.	a
Item 1	3	Item 8	4	Item 15	3	Item 22	3
Item 2	3	Item 9	3	Item 16	2	Item 23	4
Item 3	3	Item 10	4	Item 17	4	Item 24	3
Item 4	4	Item 11	3	Item 18	4	Item 25	2
Item 5	3	Item 12	1	Item 19	4	Item 26	3
Item 6	2	Item 13	3	Item 20	4	Item 27	3
Item 7	2	Item 14	3	Item 21	3	Item 28	4

country and each cycle is split into two subsamples (see section 4.2.5). The corresponding sample sizes are displayed in column 3 and 4 of Table 4.2. The integer discrimination parameters are obtained based on subsample 1. These estimated values are displayed in Table 4.3. The OPLM is estimated with subsample 2. Although the model fit test is significant, $R_{1c}(99) = 255.411$, $p < 0.0001$, the ratio between the fit statistic and the number of degrees of freedom (approx. 255/99) indicates that the fit of this model is substantially better than the fit of the PCM (approx. 2,233/231). This improvement in fit can also be seen in the item fit plot of item 17 and 25 (see Figure 4.3).

Different models in different countries

In order to investigate DIF between countries, the OPLMs are also estimated for each country, separately. The corresponding fit statistics, each time based on subsample 2, are displayed in column 5 and 6 of Table 4.2. The fit statistics demonstrate that separate models for each country does improve the fit substantially. Another way to illustrate DIF, is by comparing $\pi_{i|x_+}$ and $P_{i|x_+}$, where $\pi_{i|x_+}$ is based on item parameters obtained from a different country. For Canada and Mexico 2006, two examples are given in Figure 4.4.



(a) Item 17 (b) Item 25

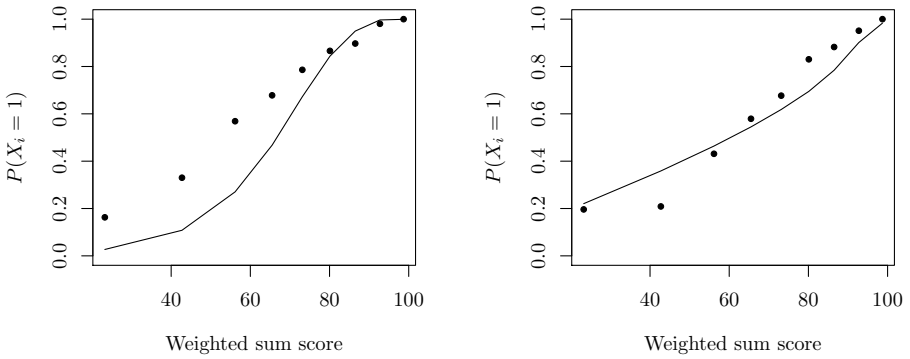
Figure 4.3: $P_{i|x_+}$ (dots), and $\pi_{i|x_+}$ (line) under the OPLM for both countries and both cycles.

The expected proportions for Canada 2006 are estimated based on the item parameters that are obtained from the OPLM for Mexico 2006, while the observed proportions are obtained from the data of Canada 2006. It can be seen that for the students in Canada in 2006 item 17 is relatively more easy, and item 21 has a larger discriminative power, compared to students in Mexico.

4.4.2 Incomplete design

In Section 4.2.2, it is suggested that the primary role of the latent variable model is to describe the expected score distribution for the missing data. This is demonstrated in the following example, based on data set 1. For each student, the responses on all 28 items are available. Now both subsamples are divided into two groups. For the first group, only the responses on item 1 to 18 are taken, i.e., item 19 to 28 are considered as missing, while for the second group only the responses on item 11 to 28 are taken. A graphical representation of the incomplete design is given in Figure 4.5.

In order to compute the plausible score distribution, the steps as described in Sections 4.2.5 and 4.2.6 are performed. The distributions of plausible scores



(a) Item 17 (b) Item 21

Figure 4.4: $\pi_{i|x_+}$ (line) and $P_{i|x_+}$ (dots), where $\pi_{i|x_+}$ is based on parameters obtained from Mexico 2006, and where $P_{i|x_+}$ is based on the data of Canada 2006.

of both countries, based on separately estimated OPLMs in each country, are displayed in Figure 4.6a. The plausible scores are displayed in grey. The observed total score on the all 28 items, is displayed in black. It can easily be seen that the plausible scores, based on 18 out of 28 items, provide a good estimation for the the sum score on all 28 item. We call this example 1.

In order to demonstrate that the estimation of plausible scores is quite robust against model misspecification, the plausible scores are also computed based on the item parameters obtained from a PCM that was fitted on the complete data of Canada and Mexico separately (example 2), and on the item parameters obtained from a PCM that was fitted on the complete data of Canada and Mexico in both 2003 and 2006 together (example 3). Although these models do not fit, Figures 4.6b and 4.6c, respectively, display that the missing data can still be estimated accurately.

4.4.3 A large data example

In this final example, the impact of DIF on the average sum score per country and the corresponding league table is illustrated. For each country, separate PCMs and OPLMs were fitted. The R_{1c}/df-ratios of the PCMs and OPLMs

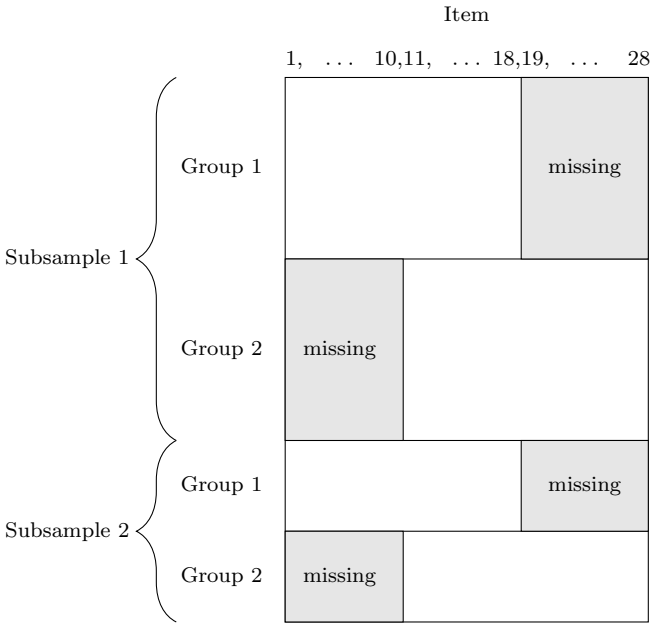
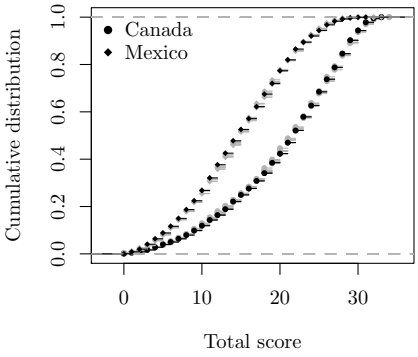


Figure 4.5: Graphical representation of the incomplete design.

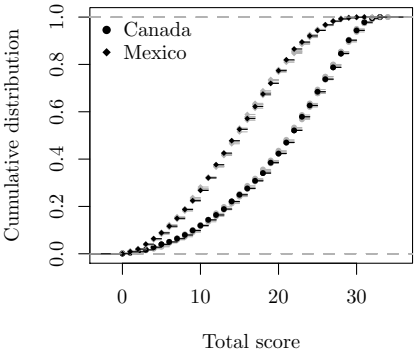
are displayed in Figure 4.7. From there it can be seen that for the PCMs only two countries have an R_{1c}/df smaller than 2, while for the OPLMs all R_{1c}/df are smaller than 2. For this example, we took the OPLM for each country.

Then, for each student, five plausible values were drawn, and these were transformed to the plausible scores metric. Since five of the items are partial credit items with three categories, the scale reaches from 0 to 25. Finally, standard errors were computed according to the regular procedure with student and replicate weights (see OECD, 2009b), Chapter 2-4).

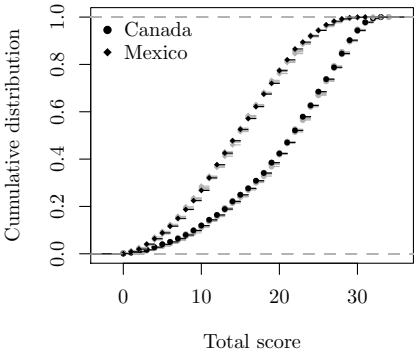
The original PISA scores (OECD, 2007) and the average plausible scores per country are depicted in Figure 4.8. The correlation between the two is 0.991. There are differences in ranks, but these are small. To illustrate the significance of the differences, the 95% confidence levels of the ranks are simulated as follows. For each country, a score was sampled from a normal distribution with mean equal to the country's mean score, and standard deviation equal to the standard error of the country's mean estimate. With these sampled scores, the rank of the countries is determined. Next, this



(a) Example 1



(b) Example 2



(c) Example 3

Figure 4.6: Distribution of observed (black) and plausible (grey) scores.

sampling scheme is repeated 10,000 times. Consequently, each country obtains 10,000 rank estimates. Finally, per country the 2.5th and 97.5th percentile of the simulated ranks are taken as a 95% certainty estimate. The resulting intervals for the original PISA scores, as well as for the plausible scores computed in this example, are displayed in Table 4.4 and Figure 4.9.

It can be seen that for the majority of countries the interval is slightly larger when based on plausible scores. This is because these scores are based on 20

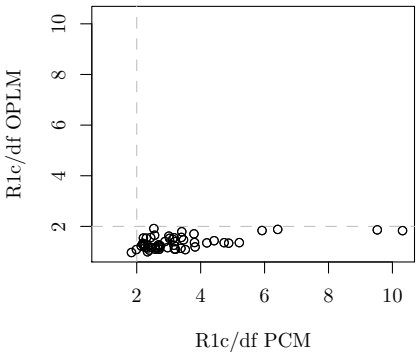


Figure 4.7: R_{1c}/df-ratios of the PCMs and OPLMs estimated per country.

items instead of 28 items, and because the plausible scores contain an additional source of uncertainty. Besides the uncertainty caused by sampling students, and the estimation error of the parameters based on the observed data, the plausible scores also contain measurement error in the sampled responses for the missing data. This latter source of uncertainty, however, is relatively small compared to the first two.

The most remarkable observation is that besides one country all other intervals do overlap. Only for Macao-China the rank based on the plausible scores on the 20 items is significantly higher compared to the PISA ranking. This overlap illustrates that taking (non-)uniform DIF into account does not affect the rank order of countries significantly.

4.5 Discussion

The aim of this paper was to provide a statistical method for educational surveys that takes DIF into account, but at the same time provides final scores that are comparable between countries. The results make clear that the model fit improves substantially if country specific item parameters are included in the method. The final league table, however, is nearly the same as the PISA league table that is based on an international calibration. This illustrates that

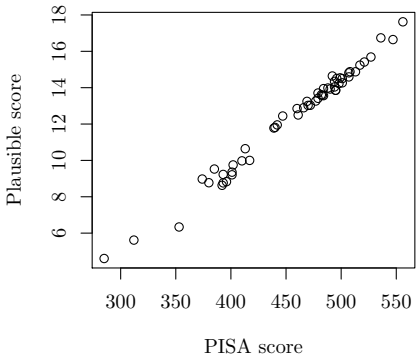


Figure 4.8: Mean scores per country based on the PISA's Reading Literacy data from 2006. On the x-axis the original PISA scores, on the y-axis the plausible scores on the subset of 20 items.

the current PISA methodology is quite robust against (non-)uniform DIF.

DIF is not the only aspect in which the PISA methodology has been criticized. Other critiques where about the multilevel and multidimensional structure of the data (Goldstein, 2004), or about the effect of item positions on item parameters (Le, 2007). These papers suggest to extend the model with additional parameters. However, when these parameters are added, then the person parameters are for the same reason as explained in this paper not directly comparable between countries. Therefore, it should be made clear how these extended models provide summary statistics that permit inter-country comparisons, something which is not done in the papers mentioned above.

Some other critiques on PISA are about the multidimensional structure of the constructs. In every PISA cycle, one domain (i.e., Math, Reading, or Science) is chosen to be the major domain. For major domains additional items are administered, and the scores on the sub domains are reported on different scales. This implies that every domain has a multidimensional structure, otherwise the breakdown into subscales would be redundant. Does this imply that every domain should be modeled with a multidimensional

model? The answer to this question is not necessarily *yes*. An alternative to the complex technique and reports of a multidimensional models could be an analysis with the following two steps. First, as described above, fit a unidimensional model in order to solve the missing data problem. Then, perform a secondary *profile analysis* (Verhelst, 2012) in order to test whether the country's profile of observed subscores is significantly different compared to the profile of subscores according to the model. This might, for instance, make clear that the country stays behind on a particular subset of the domain.

In reaction to an earlier paper of Kreiner (2011), Adams (2011) pointed out which factors need to be taken into account if someone is considering an alternative, more complicated model:

1. The need to comprehensively cover the constructs;
2. The need for analytic techniques that work simultaneously for more than 50 countries;
3. The need to integrate with appropriate sampling methodologies;
4. The requirement to provide a database that is accessible to and usable by secondary data analysts;
5. The need to support the construction of described proficiency scales;
6. The need to permit inter-country comparison of overall levels of performance;
7. The need to support the construction of scales that retain their meaning over time.

We think that our approach fits the first six requirements. The seventh, we return to later on.

In this paper, we took the sum score as a summary statistic of $\mathbf{X}$, but also another statistic could have been chosen. If we define the construct as a large set of items, then, as explained in Section 4.2.3 the data are comparable, and so are all kinds of summary statistics. Which summary statistics should be taken is a question that policy makers should answer. Statistical techniques can provide suggestions about statistics that could display interesting information, however,

there is no empirical ground for considering some statistics to be *wrong*. If we can estimate every summary statistic we like without bias from complete data, we only need to satisfy ourselves that we obtain unbiased estimates from incomplete data as well.

A question that is not considered in this paper, is which items should and should not be included in an international survey. This question is a topic of debate between the participating countries. The aim of PISA is to measure whether the knowledge and skills of 15-year old students is such that they are *prepared for life*. It could be that different countries have different (kinds of) items in mind when pondering this question. In order to find a compromise, there are mainly two options: define the greatest common divisor, or define a broader set that also contains items that fit the ideas of some but not all countries. The latter approach might cause DIF, but that, as explained in this paper, does not cause methodological problems. Instead, this diversity in the item set has the opportunity to display the diversity among countries. One way or the other, one central requirement for the method proposed in this paper is that comparisons are based on one set of items that is administered in every country. Therefore, the participating countries in PISA should agree whether responses to the final set of items does reflect preparedness for life. This set of items defines the validity of the survey instrument.

At this stage, we come back to the seventh requirement that was mentioned above, i.e., 'the need to support the construction of scales that retain their meaning over time'. Comparisons between time points, e.g., 2003 and 2006, could be of particular interest for national policy makers. To turn back to the example of the Slovak Republic that was given in Section 4.1.4, it is likely that the findings in 2003 are a reason for the government of the Slovak Republic to put extra effort in skills related to the 'uncertainty' scale. If they would do so, then it is to be expected that the differences between the three subscales becomes smaller. This would imply that, for a randomly chosen 15-year old student in the Slovak Republic, the items related to the 'uncertainty' scale will become relatively more easy. A recent report of the OECD (2012) shows that countries do react on survey outcomes. The majority of participating countries report that PISA has become an instrument for benchmarking student performance, and that PISA has had an influence on policy reform. Here, we want to emphasize that, if countries change their policy based on the survey

outcome, and if they succeed, then they actively *create* DIF between countries and within countries over time.

But not only item properties are changing over time. More general, the world around us is (rapidly) changing. Imagine, for instance, that a construct like IT literacy would have been part of the PISA survey[5]. Some items that would have been suitable in 2000, would definitely not have been suitable anymore in 2012. This implies that the content of the construct changes over time and that, in order to measure whether the skills and knowledge of 15-year-old students are sufficient for real life, the set of items should also change over (a longer period of) time. But are observed scores comparable over time if the item set changes? We think they are. Compare this case, for instance, with stock market indices like the Dow Jones Industrial Average. Because the Dow Jones Industrial Average has to reflect at each time point the *current state* of the market, the content on which the index is based changes over time (Dieterich, 2013). In the same way, if the consortium of participating countries agrees at each time point that the current set items covers the construct of interest at that particular time point, then comparison in terms observed scores are meaningful, because at each time point they reflect the construct of interest. For example, a conclusion could then be of the following structure: 'country A scores 60% of the items of 2006 correct, while they score 70% of the 2009 items correct'. Turning back to Adams's seventh requirement that an alternative method needs to support the construction of scales that retain their meaning over time, it could be concluded that the approach suggested in this paper also fits this requirement.

To conclude, if surveys have impact on policy decisions, then it is likely that item properties change over time. Moreover, the surveys could serve policy makers to change the world around us. They can compare the results of the surveys with their own benchmarks, and choose to adjust their policy. And if they succeed, survey outcomes should detect this. Therefore, surveys are part of a *dynamic system*, and what we need is methodology that does justice to these dynamics, rather than methodology that is rooted in an inherently static view on an educational system. In the this paper, we started with providing another look at DIF in international surveys. However, our approach to focus at

[5]Around 2000, it has been discussed whether this construct should be part of the PISA survey.

a large set of items also provides opportunities to study qualitative differences between countries and within countries over time. This is a topic for further research. If we more and more succeed to display all the dynamics described above, we further improve the survey instruments with respect to their main purpose: monitoring countries in a changing world.

Table 4.4: Results Reading Literacy per country: N_1 = total sample size, N_2 = sample that took reading items, UR = upper rank, LR = lower rank.

				PISA 2006				Plausible scores			
Code	Country	N_1	N_2	Score	SE	UR	LR	Score	SE	UR	LR
KOR	Korea	5171	2791	556	3.8	1	1	17.62	0.17	1	1
FIN	Finland	4710	2533	547	2.1	2	2	16.65	0.13	2	3
HKG	Hong Kong-China	4566	2465	536	2.4	3	3	16.74	0.16	2	3
CAN	Canada	22505	12129	527	2.4	4	5	15.69	0.12	4	5
NZL	New Zealand	4798	2559	521	3.0	4	6	15.41	0.17	4	6
IRL	Ireland	4572	2467	517	3.5	5	8	15.24	0.18	5	8
AUS	Australia	14081	7556	513	2.1	6	9	14.87	0.12	6	11
POL	Poland	5540	2975	508	2.8	7	11	14.88	0.14	6	12
SWE	Sweden	4420	2371	507	3.4	7	12	14.82	0.17	6	13
NLD	Netherlands	4863	2664	507	2.9	7	12	14.59	0.19	7	17
BEL	Belgium	8834	4846	501	3.0	9	16	14.27	0.18	11	21
EST	Estonia	4861	2631	501	2.9	9	16	14.50	0.16	9	17
CHE	Switzerland	12167	6579	499	3.1	10	19	14.54	0.16	8	17
JPN	Japan	5923	3196	498	3.6	10	20	14.22	0.18	12	22
TAP	Chinese Taipei	8801	4740	496	3.4	11	21	14.51	0.19	8	18
GBR	United Kingdom	13044	7045	495	2.3	13	21	13.86	0.12	18	26
DEU	Germany	4875	2704	495	4.4	11	23	13.86	0.26	15	28
DNK	Denmark	4515	2430	494	3.2	12	22	14.04	0.16	15	24
SVN	Slovenia	6589	3634	494	1.0	15	20	14.35	0.13	11	19
MAC	Macao-China	4746	2560	492	1.1	17	22	14.64	0.14	8	15
AUT	Austria	4922	2646	490	4.1	14	25	13.93	0.19	16	26
FRA	France	4673	2530	488	4.1	16	27	13.98	0.21	14	26
ISL	Iceland	3756	2009	484	1.9	22	27	13.95	0.12	17	24
NOR	Norway	4664	2507	484	3.2	21	28	13.56	0.16	22	29
CZE	Czech Republic	5927	3246	483	4.2	21	29	13.58	0.24	19	31
HUN	Hungary	4483	2401	482	3.3	22	29	13.61	0.18	20	29
LVA	Latvia	4699	2561	479	3.7	23	30	13.71	0.20	18	29
LUX	Luxembourg	4559	2446	479	1.3	25	29	13.42	0.12	25	30
HRV	Croatia	5203	2774	477	2.8	25	30	13.27	0.13	26	32
PRT	Portugal	5093	2785	472	3.6	27	33	13.03	0.19	28	34
LTU	Lithuania	4728	2544	470	3.0	29	33	13.03	0.16	29	34
ITA	Italy	21671	11636	469	2.4	30	33	13.25	0.12	27	32
SVK	Slovak Republic	4724	2550	466	3.1	30	35	12.90	0.18	29	35
ESP	Spain	19512	10516	461	2.2	33	35	12.50	0.12	34	36
GRC	Greece	4847	2609	460	4.0	32	35	12.85	0.20	30	35
TUR	Turkey	4922	2656	447	4.2	36	38	12.44	0.23	33	37
CHL	Chile	5141	2783	442	5.0	36	39	11.95	0.24	36	39
RUS	Russian Federation	5734	3076	440	4.3	36	39	11.82	0.23	37	39
ISR	Israel	4392	2362	439	4.6	36	39	11.78	0.21	37	39
THA	Thailand	6153	3342	417	2.6	40	41	10.00	0.13	41	43
URY	Uruguay	4671	2523	413	3.4	40	43	10.64	0.18	40	40
MEX	Mexico	30383	16433	410	3.1	40	43	9.97	0.15	41	43
BGR	Bulgaria	4418	2374	402	6.9	41	49	9.75	0.31	41	47
SRB	Serbia	4771	2565	401	3.5	43	47	9.35	0.16	43	48
JOR	Jordan	6433	3494	401	3.3	43	47	9.21	0.14	44	49
ROU	Romania	5102	2733	396	4.7	43	50	8.83	0.29	45	52
IDN	Indonesia	10485	5649	393	5.9	43	51	8.76	0.27	46	52
BRA	Brazil	8981	4890	393	3.7	45	50	9.23	0.15	44	49
MNE	Montenegro	4436	2369	392	1.2	46	50	8.63	0.11	49	52
COL	Colombia	4179	2259	385	5.1	47	52	9.52	0.21	42	47
TUN	Tunisia	4534	2422	380	4.0	49	52	8.77	0.20	47	52
ARG	Argentina	4068	2207	374	7.2	50	52	8.97	0.29	44	52
AZE	Azerbaijan	5184	2784	353	3.1	53	53	6.34	0.13	53	53
QAT	Qatar	6061	3281	312	1.2	54	54	5.62	0.10	54	54
KGZ	Kyrgyzstan	5313	2870	285	3.5	55	55	4.60	0.13	55	55

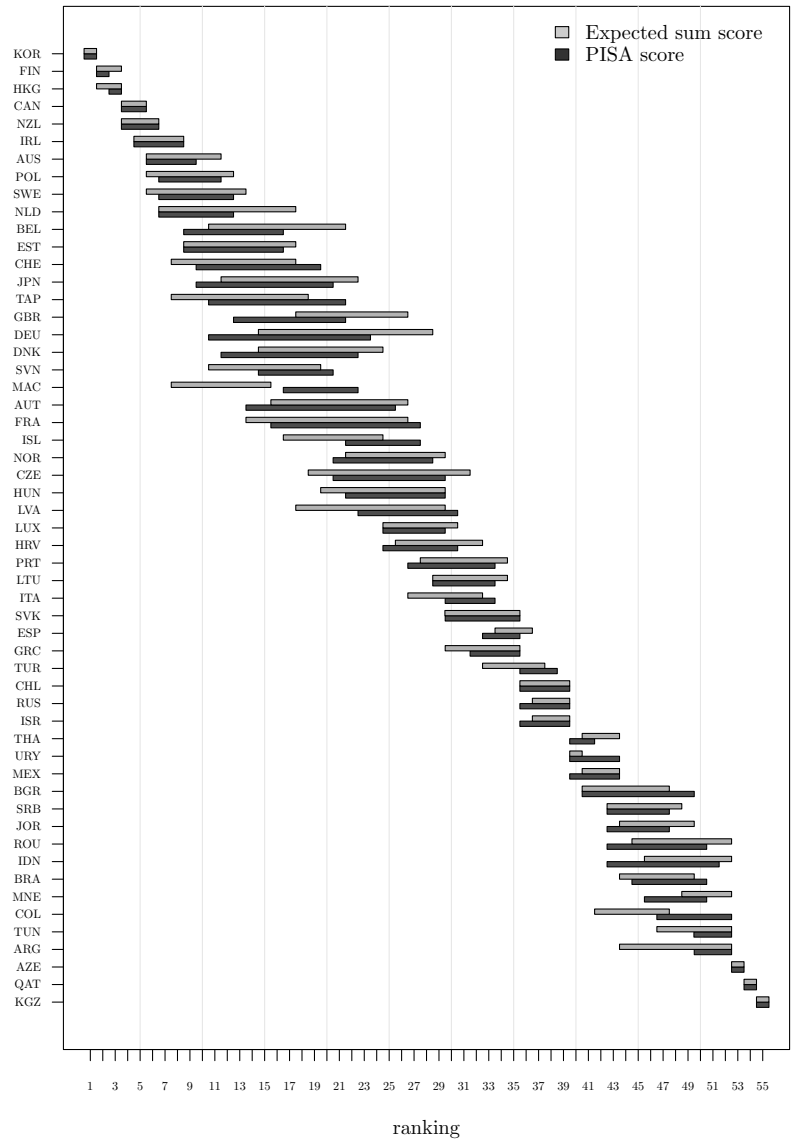


Figure 4.9: 95% rank intervals per country.