JOURNAL OF APPLIED MEASUREMENT, *8*(3), 305-322
Copyright© 2007

The Influence of Equating Methodology on Reported Trends in PISA

Eveline Gebhardt
Australian Council for Educational Research

Raymond J. Adams
Australian Council for Educational Research
University of Melbourne

In 2005 PISA published trend indicators that compared the results of PISA 2000 and PISA 2003. In this paper we explore the extent to which the outcomes of these trend analyses are sensitive to the choice of test equating methodologies, the choice of regression models and the choice of linking items. To establish trends PISA equated its 2000 and 2003 tests using a methodology based on Rasch Modelling that involved estimating linear transformations that mapped 2003 Rasch-scaled scores to the previously established PISA 2000 Rasch-scaled scores. In this paper we compare the outcomes of this approach with an alternative, which involves the joint Rasch scaling of the PISA 2000 and PISA 2003 data separately for each country. Note that under this approach the item parameters are estimated separately for each country, whereas the linear transformation approach used a common set of item parameter estimates for all countries. Further, as its primary trend indicators, PISA reported changes in mean scores between 2000 and 2003. These means are not adjusted for changes in the background characteristics of the PISA 2000 and PISA 2003 samples – that is, they are marginal rather than conditional means. The use of conditional rather than marginal means results in some differing conclusions regarding trends at both the country and within-country level

Requests for reprints should be sent to Eveline Gebhardt, Australian Council for Educational Research, 19 Prospect Hill Rd., Camberwell, VIC 3124, Australia, e-mail: gebhardt@acer.edu.au.

Introduction

Comparing and ranking country performances are not the only goals of the PISA study. Another main goal is to estimate trends within countries over time. To enable comparisons across cycles, the PISA 2000 and PISA 2003 assessments of mathematics, reading and science were linked assessments. That is, the sets of items used to assess each of mathematics, reading and science in PISA 2000 and the sets of items used to assess each of mathematics, reading and science in PISA 2003 included a subset of items common to both sets. These common items are referred to as link items. The number of link items within each domain was 20 for mathematics, 28 for reading and 25 for science.

In the case of mathematics a decision was made to produce a new scale for PISA 2003 and not to report overall trends because the combined mathematics domain of PISA 2003 included subscales that were not included in PISA 2000 (OECD, 2004).

The procedures and models used to scale PISA data are fully described in Adams and Carstensen (2002) and Adams, Wu, and Carstensen, (2006) and will not be discussed in this article. The software ACER ConQuest (Wu, Adams and Wilson, 1998) is used for PISA scaling. The steps involved in the original linking of PISA 2000 and PISA 2003 reading and science scales are described in detail in OECD (2005b) and can be summarized as follows (see also Figure 1).

Step 1. The PISA 2000 data from each of the OECD countries were re-scaled with full conditioning (see Mislevy and Sheehan, 1987; Adams et al. 2006) and with link items anchored at their international PISA 2003 values.[1]

Step 2. The mean and standard deviation of each literacy domain was then calculated for the combined 2000 data set of 25 OECD countries. Each country was given equal weight.

Step 3. The mean and standard deviations computed in Step 2 were then compared with the matching means and standard deviations from the original PISA 2000 scaling.[2] Linear transformations that mapped the PISA 2003 based scores to scores that would yield a mean and standard deviation equal to the PISA 2000 results were then computed.

Step 4. The linear transformation from step 3 was applied to the PISA 2003 scales to provide an equating to the PISA 2000 scales.

From the perspectives of participating countries, information about trends in literacy outcomes over time are amongst the most important reasons for participating in PISA. Changes in either the ranking of countries over time or in the absolute level of performance in a country are seen as fundamental indicators as to success or otherwise of education systems (OECD, 2004). At finer grained levels participating economies are also concerned with other trends such as changes in performance variation, gender effects, socio-economic status effects and the like. This paper is concerned with the potential impact that *technical* choices in the equating methodology might have on trend results that are of interest to participants.

Three particular issues in trend methodology are examined in this paper. The first issue is concerned with how the item parameters are estimated. Current practice in PISA is to estimate a single set of *international* item parameters and to apply them to the proficiency estimation of students in each country, the single set of Rasch item parameter estimates are obtained by calibrating the items on a representative sub-sample of 30 000 students drawn from the 30 participating OECD countries (Adams 2002; OECD 2005b). This approach is undertaken because it is consistent with the intention of measuring all students, regardless of their country, on a common scale.

A potential weakness of this approach is that it ignores item-by-country interactions in the behavior of the items. Item-by-country interactions,

[1] The international item parameter estimates for PISA 2003 are published in OECD (2005b).

[2] The international item parameter estimates for PISA 2000 are published in Adams and Wu (2002).

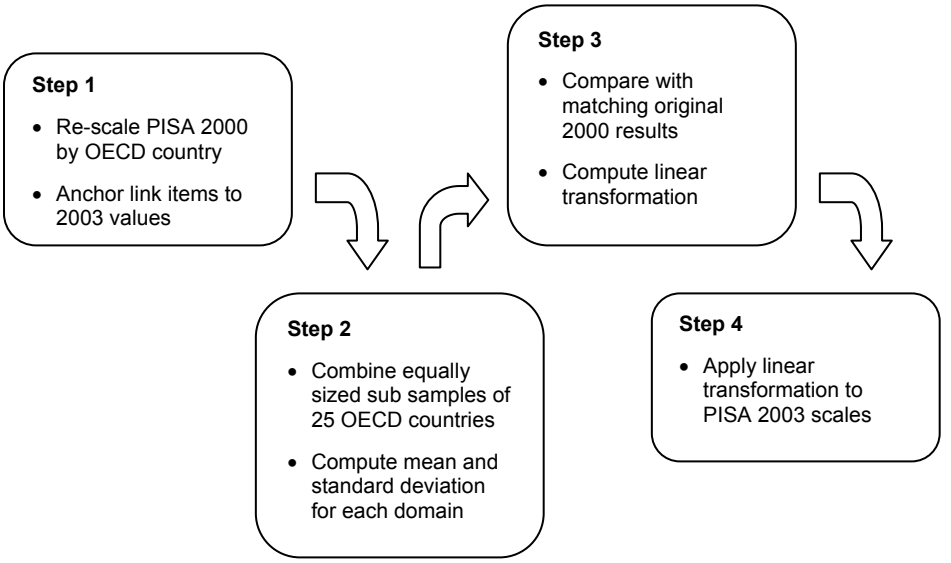

Figure 1. Steps involved in original linking of PISA 2000 and PISA 2003.

a form of country DIF, are commonly observed in cross-national studies (Adams and Carstensen, 2002; Adams, Wu and Macaskill, 1998; Mullis and Martin, 1998) and the magnitude of these interactions influences the validity of cross-country comparisons.

Rather than estimating trends with a common set of item parameters an alternative is to estimate a unique set of parameters for each country with the scales made comparable by setting a common mean for the set of items that were used in all countries. The trend estimates for each country would then be less comparable across countries, but within countries, they would be based upon more appropriate item parameter estimates.

The second issue relates to how the PISA 2000 and PISA 2003 scales are linked. As mentioned above and described in OECD (2005b) the transformation from the 2003 to 2000 scale is set by re-scaling the 2000 data sets with items anchored at their 2003 values and then computing the transformation that would map the estimates from the rescaling onto the same scale as the original scaling. These procedures follow those used in the US National Assessment of Educational Progress and is a form of common item equating sometimes referred to as Stocking-Lord equating (Johnson and Carlson 1992, Hedges and Vevea, 2003).

An alternative and more common approach in other settings would be to merge the two data sets so that items are concurrently calibrated (Kim and Cohen, 1998). In this case the resulting item parameter estimates and proficiency estimates are automatically located on a common scale and no linear transformations are needed for equating. Note that this approach is not feasible in the practice of studies like PISA because it necessarily involves computing trends based upon results (in this case the 2000 results) that may well be different to those that had been originally published.

A third issue relates to the extent to which trends should be adjusted to account for changes over time in the demography of the target populations and or the achieved samples. For example, suppose that for some reason two data sets from different cycles have different proportions of male and female students and further assume that this difference is due to sampling variation rather than a real change in the population proportions. Under such circumstances, controlling for the proportion of males in the calculation of trends could lead to a more interpretable (and perhaps valid) measure of the trend in that particular

country. As another example, however, imagine a country with a positive trend, not adjusted for possible changes in background characteristics of the students, with a proportion of males that is stable across cycles, but has an increase in the number of missing values for gender. If students with missing values for gender perform poorly, as typically is the case, including an indicator for these missing values in the multiple regression analysis to adjust the trend for changes in the sample would lead to an overestimation of the real trend. Here the unadjusted trend is the preferred trend indicator.

In this paper unadjusted trends are called *marginal* and adjusted trends *conditional*.

Two alternative trend estimates were computed and compared to the original results and to each other (see Table 1). Both alternative trends were based on national item parameters and joint scaling of the PISA 2000 and PISA 2003 data sets. One of the trend estimates is the marginal difference (unadjusted) between cycles for each country, the other is the conditional difference controlling for student background variables that were collected in both cycles.

Table 1

Variables in trend methodology

		Trend Indicator	
Item parameters	Equating method	Unadjusted: Marginal trend	Adjusted: Conditional trend
International item parameters	Joint calibration	—	—
	Linear transformation	(1) Original PISA	—
National item parameters	Joint calibration	(2) This paper	(3) This paper
	Linear transformation	—	—

Methodology

Scaling for this paper was undertaken in a similar way to the original PISA analysis. Because such a scaling involves numerous steps and is somewhat complex, full details are not given here. The interested reader is referred to the PISA technical reports and data analysis manual (Adams, 2002; OECD, 2005a; OECD 2005b). Student sampling weights were used, booklet corrections were applied, plausible values were drawn and balanced repeated replication (BRR) method (Judkins, 1990) was used to derive unbiased standard errors. Special education students (that were assessed with special booklets) were excluded from item calibration and included when generating student scores.

Three steps had to be undertaken to estimate the two alternative trends. First, national item parameters were estimated in uni-dimensional models, using combined national data sets of PISA 2000 and PISA 2003 with only students that responded to items in that domain (excluding students that responded to the special education booklet) and both link and non-link items. In this model, a variable for booklet was used as a facet (Linacre, 1994), with a maximum (depending on which domain) of nine booklets from PISA 2000 and a maximum of 13 booklets from PISA 2003. Booklet had to be added as a facet because variation in booklet means was sometimes larger than expected (Adams and Carstensen, 2002). The inclusion of an item facet removed, in part, the influence of item-position-within booklet effect on the item parameter estimates.

Second, a three-dimensional Rasch model was run with mathematics, reading and science as domains to draw plausible values anchoring item parameters to the values that were estimated in the first step. All students (including special education) and all link and non-link items were included in the model. Deviation contrast coding was used for the coding of the test booklet, a dummy coding was used for PISA cycle, and these variables were used as latent regression variables. Latent regression variables had to be used to

make the three-dimensional model mathematically equivalent to the uni-dimensional models with booklets as facets.[3] Therefore, booklet facets were transformed into deviation contrasts and as such included in the multi-dimensional conditioning model. Deviation booklet contrasts were designed so that the reference group is the full group of students that did respond to items in a domain. Including all these students was important because this reference group is used as the intercept when imputing abilities for students that did not respond to items in that domain. Since the variation in booklet means was larger than expected, using a reference group with only students responding to one particular booklet can lead to over- or underestimation of students' abilities that did not respond to items in a domain. Alongside these booklet and cycle indicators, common background variables of both cycles were included as regression variables as well. These common background variables were sex, highest socio-economic status of parents (HISEI[4]), age, language at home, school mean in mathematics, reading and science, mother's occupation, and father's occupation and their indicators for missing values.

Finally, single and multiple regression analyses were run in SPSS (using student sampling weights and BRR replication method) to compute both the marginal and conditional trends and their probabilities. In the single regression analysis, a dummy for cycle (0=PISA 2000, 1=PISA 2003) was the only independent variable. The regression coefficient for this dummy was used as the marginal trend for a domain within a country.

In the multiple regression analysis, the following variables were included as independent variables:

- age (in months and missing replaced by mean age within cycle);
- missing age (1=missing, 0=not missing);
- sex (1=girl, 0=boy or missing);
- missing sex (1=missing, 0=not missing);
- HISEI (scale from 16 to 90 and missing replaced by mean HISEI within cycle);
- missing HISEI (1=missing, 0=not missing);
- language at home (0=test language or missing, 1=other language);
- missing language at home (1=missing, 0=not missing); and,
- cycle (0=PISA 2000, 1=PISA 2003).

This set of variables was chosen because they are the key non-malleable variables that are available in the PISA database. That is, they are variables that are not amenable to influence through educational policy and practice and as such it is commonly argued that comparisons across countries and over time might be more valid if they were adjusted for student variation in these variables. The (partial) regression coefficient of the cycle dummy was used as the conditional trend for a domain within a country.

Note that estimates of linking errors (Monseur and Berezner, 2007) were not included in the standard errors reported in this paper. One reason is that no consensus has been reached about this issue. Another, more practical reason, is that all countries have different item parameters, unlike in the original analysis, which means that the linking error has to be estimated for each country separately. To complicate the estimation further, some countries have nationally deleted link items due to mistranslations.

Results

Twenty-eight countries were included in the analysis for this study. The participating countries, the number of weighted and unweighted students and the original mean performances in PISA 2000 and PISA 2003 are listed in Table 2.

The list of countries consists of 27 of the 28 OECD countries that participated in both cycles and the Russian Federation. Greece, an OECD country that participated in both cycles, was excluded on the advice of the PISA sampling referee

[3] Note that the ConQuest cannot be easily configured to permit different facet effects for each dimension.

[4] The PISA variable HISEI is an index of the socio-economic status of the student's family it is derived following the procedures of Ganzeboom, de Graaf and Treiman (1992).

Table 2

Participating countries, number of weighted and unweighted students and original mean performance in PISA 2000 and PISA 2003

	Unweighted *N*		Weighted *N*		Reading–P2000		Reading–P2003		Science–P2000		Science–P2003	
	P2000	P2003	P2000	P2003	Mean	*SE*	Mean	*SE*	Mean	*SE*	Mean	*SE*
Australia	5176	12551	229152	235591	528	(3.5)	526	(2.2)	528	(3.5)	529	(2.1)
Austria	4745	4597	71547	85931	507	(2.4)	498	(3.8)	519	(2.5)	494	(3.6)
Belgium	6670	8796	110095	111831	507	(3.6)	509	(2.6)	496	(4.3)	513	(2.4)
Canada	29687	27953	348481	330436	534	(1.6)	531	(1.8)	529	(1.6)	527	(2.0)
Czech Republic	5365	6320	125639	121183	492	(2.4)	489	(3.6)	511	(2.4)	524	(3.3)
Denmark	4235	4218	47786	51741	497	(2.4)	501	(2.8)	481	(2.8)	487	(3.0)
Finland	4864	5796	62826	57884	546	(2.6)	540	(1.7)	538	(2.5)	550	(1.9)
France	4673	4300	730494	734579	505	(2.7)	499	(2.6)	500	(3.2)	518	(3.1)
Germany	5073	4660	826816	884358	484	(2.5)	495	(3.3)	487	(2.4)	505	(3.6)
Hungary	4887	4765	107460	107044	480	(4.0)	480	(2.5)	496	(4.2)	498	(2.7)
Iceland	3372	3350	3869	3928	507	(1.5)	494	(1.4)	496	(2.2)	493	(1.5)
Ireland	3854	3880	56209	54850	527	(3.2)	520	(2.5)	513	(3.2)	511	(2.6)
Italy	4984	11639	510792	481521	487	(2.9)	471	(3.0)	478	(3.1)	479	(3.2)
Japan	5256	4707	1446596	1240054	522	(5.2)	507	(3.7)	550	(5.5)	536	(4.2)
Korea	4982	5444	579109	533504	525	(2.4)	540	(3.1)	552	(2.7)	541	(3.5)
Luxembourg	3404	3923	4138	4080	441	(1.6)	479	(1.2)	443	(2.3)	483	(1.4)
Mexico	4600	29983	960011	1071650	422	(3.3)	394	(4.2)	422	(3.2)	394	(3.7)
Netherlands	2503	3992	157327	184943	532	(3.4)	516	(2.9)	529	(4.0)	529	(3.2)
New Zealand	3667	4511	46757	48638	529	(2.8)	525	(2.7)	528	(2.4)	525	(2.5)
Norway	4147	4064	49579	52816	505	(2.8)	497	(2.7)	500	(2.7)	482	(3.0)
Poland	3654	4383	542005	534900	479	(4.5)	495	(2.9)	483	(5.1)	494	(3.0)
Portugal	4585	4608	99998	96857	470	(4.5)	473	(3.7)	459	(4.0)	472	(3.4)
Russia	6701	5974	1968131	2153373	462	(4.2)	439	(3.9)	460	(4.7)	483	(4.2)
Spain	6214	10791	399055	344372	493	(2.7)	478	(2.5)	491	(3.0)	479	(2.6)
Sweden	4416	4624	94338	107104	516	(2.2)	515	(2.5)	512	(2.5)	510	(2.8)
Switzerland	6100	8420	72010	86491	494	(4.2)	505	(3.1)	496	(4.4)	514	(3.6)
United Kingdom	9340	9535	643041	698579	523	(2.6)	511	(2.4)	532	(2.7)	523	(2.7)
United States	3846	5456	3121874	3147089	504	(7.0)	495	(3.0)	499	(7.3)	494	(3.0)

(K. F. Rust, personal communication, 8 March 2006) due to observed inconsistencies in students' weights. The Netherlands and United Kingdom were not included in the official PISA reports, because they had unacceptable non-response rates in one of the cycles. They are included here, but the results should be interpreted with caution.

Three sets of results were compared: the original results for trends and the results from two alternative methods of equating. There was some difference between the original results presented here and the results that were previously published (OECD, 2001), because the 2003 data was rescaled using deviation contrast coding for booklets instead of simple contrast coding. Results from the first alternative method are called *marginal trends*. They were the differences in means between cycles for each country. The second alternative results, called *conditional trends*, were the differences between cycle means after controlling for the common background variables sex, age, HISEI, and language spoken at home and their respective dummies for missing values.

Table 3 presents the significance of the differences for each country between PISA 2000 and PISA 2003 in reading and science (see Appendix A and B for the regression coefficients, standard errors and standard normal scores). First, the differences between cycles of the *original* results are presented, then the *marginal* differences and finally the *conditional* differences. Figure 2 and Figure 3 are a graphical representation of the trends in standard normal scores. In these figures, each country has three marks, one for each trend estimate. The square mark represents the original trend, the diamond the marginal trend and the

triangle the conditional trend. Each trend estimate is the difference in country mean between 2000 and 2003, divided by the standard error of the difference to derive a standardised score. Positive numbers indicate an increase in country performance between 2000 and 2003, negative numbers a decrease. The horizontal lines are reference lines for testing significance of these changes over time. Trend estimates in between the thin lines are not significant ($p > .05$). In other words, the country's performance did not change over time. Trend estimates above and below the thick lines are significant at the .01 level, above and below the thin horizontal lines at the .05 level. The focus of these figures is the distance between the three marks for each country and the possible corresponding change in significance. In order to simplify the graphical presentations the outlying trend estimates of Luxemburg were excluded from the figures.

Original versus Marginal trends

For the original calculation of trends, items parameters were estimated using an international calibration sample, separately for PISA 2000 and PISA 2003. In contrast, the marginal trends were

Table 3

A comparison of the significance of trends between three alternative methods for equating

	Reading			Science		
	Original	Marginal	Conditional	Original	Marginal	Conditional
Australia	0	0	0	0	0	- -
Austria	-	- - -	-	- - -	- - -	- - -
Belgium	0	0	0	+++	+++	+++
Canada	0	0	+++	0	0	++
Czech Republic	0	0	0	+++	++	0
Denmark	0	0	0	0	0	0
Finland	- -	0	0	+++	0	0
France	0	0	0	+++	+++	+++
Germany	+++	0	+++	+++	+++	+++
Hungary	0	0	0	0	0	0
Iceland	- - -	- - -	- - -	0	0	0
Ireland	-	0	0	0	0	0
Italy	- - -	- -	- - -	0	0	0
Japan	- -	0	- - -	- -	- -	- - -
Korea	+++	+++	+++	- -	- - -	- - -
Luxembourg	+++	+++	+++	+++	+++	+++
Mexico	- - -	- - -	- - -	- - -	- - -	- - -
Netherlands	- - -	- - -	- - -	0	0	0
New Zealand	0	- -	- - -	0	-	- - -
Norway	- -	- - -	- - -	- - -	- - -	- - -
Poland	+++	+++	+++	+	+++	+++
Portugal	0	+	0	++	++	+
Russian Federation	- - -	- - -	- - -	+++	+++	++
Spain	- - -	- - -	- - -	- - -	- -	- - -
Sweden	0	- - -	- - -	0	0	0
Switzerland	++	+	0	+++	+++	+++
United Kingdom	- - -	- - -	- - -	- -	- - -	- - -
United States	0	0	- - -	0	0	- -

Note: Significance level:	2003 better than 2000:	2003 worse than 2000:
$p > .10$	0	0
$p < .10$	+	-
$p < .05$	++	- -
$p < .01$	+++	- - -

based on national item parameters and a combined PISA 2000 and PISA 2003 data set. These different methods resulted for some countries in substantial differences in trends. For example, the original results suggested that Swedish students from PISA 2003 performed as well in reading as Swedish students from PISA 2000 (see Figure 2). However, when using joint calibration and national item parameters, students from PISA 2003 performed significantly worse than students from PISA 2000 (p <0.01). This difference between original trend and marginal trend was far less for some other countries. A few variables could explain the variation of these differences across countries.

One hundred and twenty-nine items were administered to all students in PISA 2000. Of these

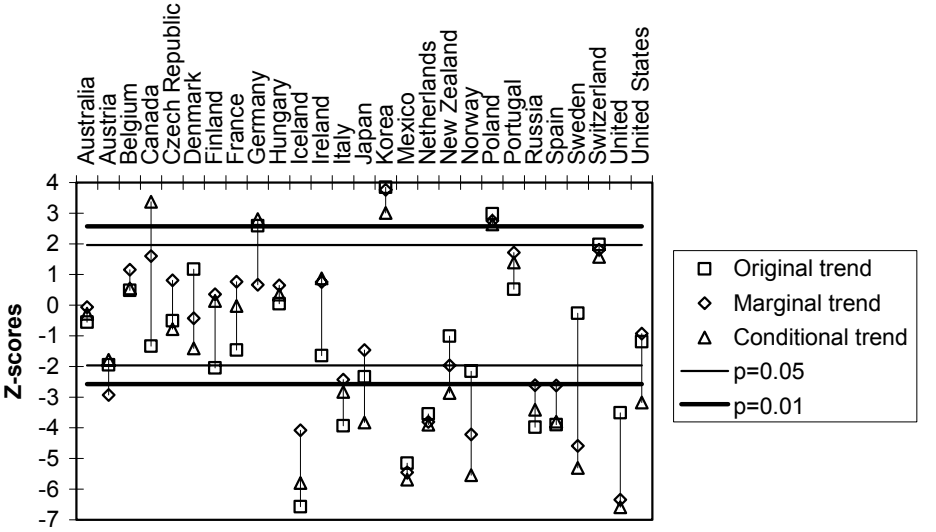

Figure 2. Three alternative trends between PISA 2000 and PISA 2003 in reading by country

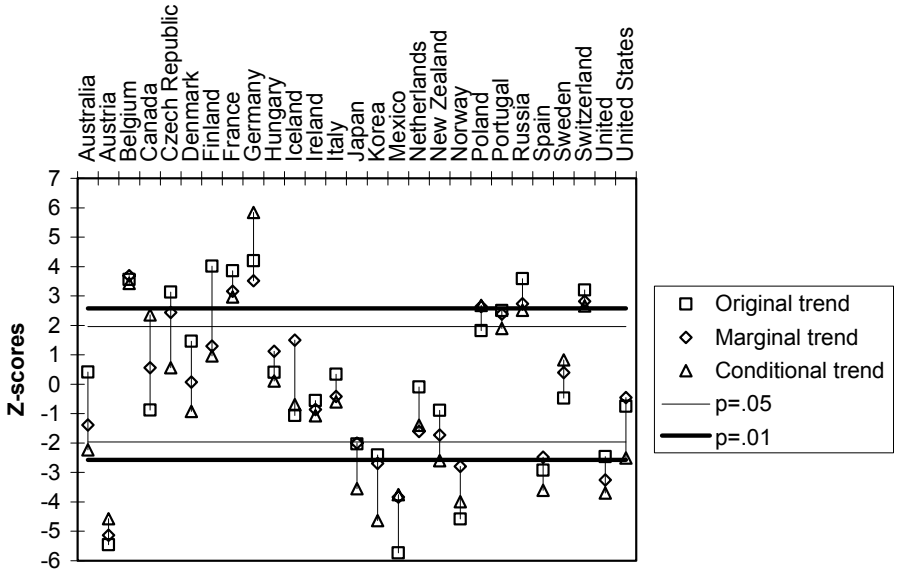

Figure 3. Three alternative trends between PISA 2000 and PISA 2003 in science by country

129 items, 28 were administered again in 2003 (no new items were added in 2003). The average international difficulty of these 28 link items was –0.03 (in logits), while the 101 unique PISA 2000 items had an average difficulty of 0.01 (in logits). Therefore, the link items were slightly *easier* than the non-link items (0.04 of a logit) when using international item parameters. The difference between the national average difficulty of link items and non-link items is called the *relative difficulty of link items* and varies across countries. In other words, this is a form of test-by-country interaction. These average reading item difficulties and the relative difficulty of the set of link items are displayed in Table 4, followed by an explanation of the figures in the table.

The second column in Table 4 is the average difficulty of unique PISA 2000 items. The third column is the average difficulty of the link items. The national relative difficulty of link items in the next column is the difference between those two columns. Positive national relative difficulties of link items indicate that the link items were easier than unique PISA 2000 items. For example, the average difficulty of the Swedish reading link item parameters was –0.20 and the average of non-link item parameters was 0.06. Therefore, the link items were on average 0.26 of a logit easier than the non-link items in Sweden.

This national relative difficulty of link items needed to be compared with the international value, because the international parameters were used as anchors in the original scaling, as will be illustrated in the next paragraph. The international relative difficulty of link items is computed in the first two

Table 4

Computation of relative difficulty of reading link items

	Reading				
	Unique 2000	Links	National relative difficulty links	(Adjusted) International relative difficulty	Relative Difficulty Links
International 2000	0.01	–0.03		0.04	
International 2003		–0.41			
Australia	0.01	–0.04	0.05	0.04	0.01
Austria	0.02	–0.06	0.07	0.04	0.03
Belgium	–0.01	0.03	–0.04	0.04	–0.08
Canada	–0.02	0.05	–0.07	0.04	–0.11
Czech Republic	–0.01	0.04	–0.05	0.04	–0.09
Denmark	0.03	–0.12	0.15	0.04	0.11
Finland	0.01	–0.04	0.05	0.04	0.01
France	–0.02	0.07	–0.09	0.04	–0.13
Germany	0.01	–0.05	0.06	0.04	0.02
Hungary	–0.03	0.09	–0.11	0.03	–0.14
Iceland	–0.01	0.00	–0.01	0.03	–0.04
Ireland	–0.01	0.02	–0.03	0.04	–0.07
Italy	–0.03	0.09	–0.12	0.02	–0.14
Japan	–0.03	0.09	–0.12	0.04	–0.16
Korea	0.01	–0.04	0.06	0.08	–0.03
Luxembourg	–0.01	0.03	–0.04	0.04	–0.08
Mexico	–0.05	0.18	–0.23	0.04	–0.27
Netherlands	–0.01	–0.04	0.03	0.03	0.00
New Zealand	0.03	–0.10	0.13	0.04	0.09
Norway	0.02	–0.09	0.11	0.04	0.08
Poland	0.00	0.07	–0.07	0.01	–0.08
Portugal	–0.04	0.14	–0.18	0.04	–0.22
Russia	–0.04	0.11	–0.15	0.05	–0.21
Spain	–0.02	0.08	–0.10	0.05	–0.15
Sweden	0.06	–0.20	0.26	0.05	0.21
Switzerland	0.01	–0.03	0.04	0.04	0.00
United Kingdom	0.01	–0.12	0.14	0.04	0.10
United States	0.02	0.01	0.01	0.04	–0.03

rows and has a value of 0.04, indicating that the link items were 0.04 of a logit easier than the unique PISA 2000 items when using the international calibration sample. To complicate this one step further, some countries deleted some items because of mistranslations. These deletions had an effect on the international relative difficulty. Therefore, the international value was adjusted for countries with nationally deleted items (see column 5). Korea, for example, deleted four reading items, which resulted in an international relative difficulty of 0.08. The last column compares the national with the (adjusted) international value and is our final measure for *relative difficulty of link items*.

The variation in relative difficulty of link items (the last column in Table 4 and in Table 5) was related to the variation in difference between original and marginal trends. This is illustrated with the Swedish example of the trends in reading. As mentioned before, the national calibration showed that for Swedish students the reading link items were 0.26 of a logit *easier* than the unique reading PISA 2000 items. These national parameters were used for computing the marginal trend. However, when calculating the original trends, the item parameters were fixed to the international values where the link items were only 0.04 of a logit *easier* than the unique PISA 2000 items. Since the link items were the only items administered in PISA 2003, the abilities of the Swedish students were higher in the original scaling of PISA 2003, using international item parameters, than the abilities from the national scaling for the marginal trend. Therefore, the original trend in Sweden was less negative than the marginal trend.

Both PISA 2000 and PISA 2003 assessed unique as well as common science items, which makes the computation of relative difficulty of science link item somewhat more complex. The fourth column in Table 5 is the relative difficulty of science link items in PISA 2000 (column two minus column four) and the fifth column under science is the relative difficulty of link items in PISA 2003 (column three minus column four). The national relative difficulty of link items is the difference between column five and six and is listed in column seven. The next column lists the international relative difficulty for link items adjusted to take into account national deletions. For the international calibration sample, the link items were 0.10 of a logit *easier* than the unique PISA 2000 items and 0.02 of a logit *harder* than unique PISA 2003 items. These figures were opposite in Norway where the link items were 0.07 of a logit *harder* than unique PISA 2000 items and 0.04 *easier* than unique PISA 2003 items. Again, the comparisons between the national and international value give the final *relative difficulty of link items* and is listed in the last column of Table 5.

The effect of relative difficulty of science link items is illustrated for Denmark. Using the international item parameters, the link items were 0.10 of a logit easier than the unique PISA 2000 items, which was very similar to the national item parameters of Denmark (0.11 of a logit easier). In the international calibration of PISA 2003, the international students found the link items 0.02 of a logit harder than the unique PISA 2003 items, or, in other words, the unique PISA 2003 items were 0.02 of a logit *easier* than the link items. However, for the Danish students the unique PISA 2003 items were 0.17 of a logit *easier* than the link items. Therefore, using international item parameters of PISA 2003 resulted in higher abilities than using national item parameters for Danish students in 2003. The result was a more positive *original* trend than *marginal* trend (see Figure 3). Figure 4 and Figure 5 give the scatter plots for reading and science between the relative difficulty of link items and the difference between original and marginal trends.

The correlation for reading was 0.59 and for science 0.58. Luxemburg is an outlier in both plots with an extreme positive trend. If Luxemburg is deleted the correlations are 0.67 and 0.82 respectively.

Marginal versus Conditional trends

As discussed above if there are differences in the marginal and conditional trends the difference is caused by changes in the distribution of the sample across the categories of the non-malleable background variables that are included in the regression model. As we expected an examination

Table 5

Computation of relative difficulty of science link items

	Science							
	Unique 2000	Unique 2003	Links	Relative difficulty links 2000	Relative difficulty links 2003	National relative difficulty links	(Adjusted) International relative difficulty	Relative Difficulty Links
International 2000	0.08		–0.03	0.10				
International 2003		0.08	0.10		–0.02		0.13	
Australia	0.23	–0.13	–0.04	0.27	–0.10	0.36	0.13	0.23
Austria	–0.03	–0.06	0.03	–0.07	–0.09	0.03	0.13	–0.10
Belgium	0.11	–0.13	0.01	0.11	–0.14	0.25	0.13	0.12
Canada	0.04	0.07	–0.04	0.08	0.11	–0.03	0.13	–0.15
Czech Republic	0.09	–0.01	–0.03	0.12	0.02	0.10	0.13	–0.03
Denmark	0.12	–0.16	0.01	0.11	–0.17	0.28	0.13	0.15
Finland	0.13	–0.33	0.07	0.05	–0.40	0.45	0.13	0.33
France	0.16	–0.11	–0.02	0.18	–0.09	0.27	0.13	0.14
Germany	0.11	–0.25	0.05	0.08	–0.19	0.27	0.12	0.15
Hungary	–0.03	–0.05	0.03	–0.06	–0.08	0.02	0.13	–0.11
Iceland	–0.05	–0.06	0.04	–0.03	0.13	–0.16	0.05	–0.21
Ireland	0.06	–0.07	0.00	0.05	–0.07	0.13	0.13	0.00
Italy	0.17	–0.15	–0.01	0.18	–0.14	0.32	0.13	0.20
Japan	0.02	0.03	–0.02	0.04	0.05	–0.01	0.13	–0.14
Korea	–0.04	0.00	0.01	–0.05	–0.01	–0.04	0.13	–0.17
Luxembourg	0.06	–0.05	0.00	0.06	–0.05	0.11	0.13	–0.02
Mexico	–0.03	0.25	–0.08	0.05	0.32	–0.28	0.13	–0.40
Netherlands	0.10	–0.13	0.01	0.11	–0.06	0.17	0.13	0.04
New Zealand	0.08	–0.07	0.00	0.08	–0.07	0.15	0.13	0.03
Norway	–0.09	0.01	–0.02	–0.07	0.04	–0.11	0.12	–0.23
Poland	0.05	0.05	–0.04	0.09	0.09	0.00	0.13	–0.12
Portugal	0.07	0.05	–0.09	0.16	0.14	0.01	0.08	–0.07
Russia	0.29	–0.09	–0.06	0.30	–0.03	0.33	0.11	0.22
Spain	–0.01	–0.02	0.01	–0.02	–0.03	0.01	0.13	–0.11
Sweden	0.05	–0.02	–0.01	0.06	–0.01	0.07	0.13	–0.06
Switzerland	0.17	–0.19	0.01	0.17	–0.20	0.36	0.13	0.24
United Kingdom	0.17	–0.08	–0.03	0.21	–0.05	0.25	0.13	0.12
United States	0.07	0.17	–0.09	0.16	0.25	–0.10	0.13	–0.22

of the cases where the marginal and conditional trends are different leads to different preferences, on a case-by-case basis for either the marginal or conditional trend estimates.

In the following, we examine more closely those cases where the marginal trends and the conditional trends lead to different interpretations.

AUSTRALIA

The conditional trend in reading was less negative than the marginal trend.

Although the trends in science show different levels of significance in Table 3, Figure 3 shows that the actual difference was rather small. The change in *Z*-score is largely caused by a change in standard error, not in the estimate of the regression coefficient—marginal regression coefficient is –0.06 (standard error 0.035), conditional regression coefficient is –0.08 (standard error 0.044). The small difference between the marginal and conditional trends is mainly caused by an increase in age of one month (mean age is 188 months in PISA 2000 and 189 months PISA 2003) and a decrease in percentage of students that do not speak the test language at home (17 percent in PISA 2000 and nine percent in PISA 2003). If a choice has to be made between the two trends, the conditional trend is probably a slightly better estimate of the real trend.

AUSTRIA

The conditional trend in reading was less negative than the marginal trend.

In PISA 2000, Austria had 47 percent boys, a mean HISEI of 50 and six percent of students did not speak the test language at home most of

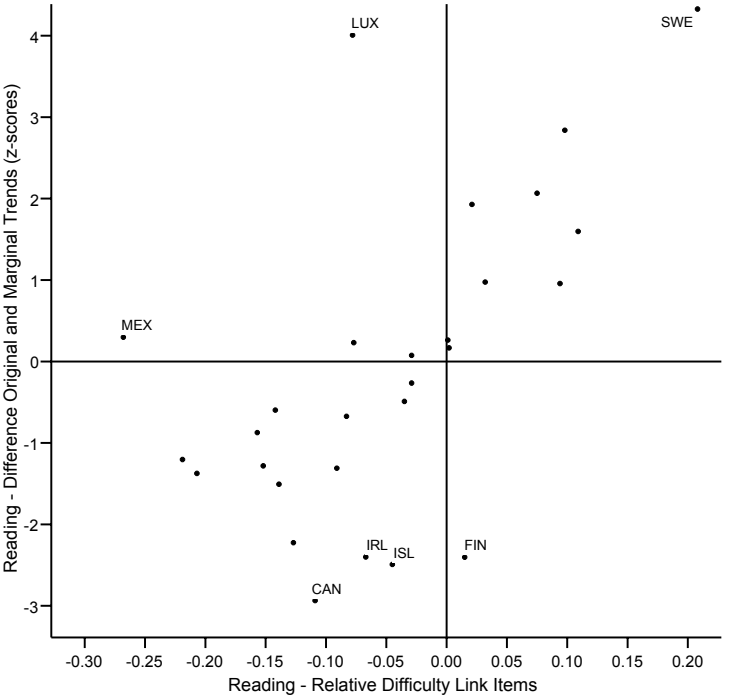

Figure 4. Scatterplot between relative difficulty of reading link items versus difference in original and marginal trends in reading

the time. In PISA 2003, these figures were 50 percent boys, mean HISEI of 47 and nine percent of students that do not speak the test language at home. The drop in HISEI seemed to be the major cause of the more negative marginal trend, which suggest that the real trend is closer to the conditional than the marginal trend.

CANADA

The conditional trends in science and reading were more positive than marginal trends.

In PISA 2003, seven to eight percent more students had missing values for the variables sex (zero percent in PISA 2000, seven percent in PISA 2003), age (zero percent in PISA 2000, eight percent in PISA 2003) and language at home (three percent in PISA 2000, 11 percent in PISA 2003) than in PISA 2000. The percentage of boys, average age and percentage of students speaking the test language at home stayed approximately the same. Especially the missing indicator for language at home had a strong negative effect on science performance. Listwise deletion resulted in a conditional trend that was very similar to the marginal trend. Since the distribution of valid responses on the background variables had not changed over time and because there was no valid reason for deleting students with missing values on these variables, it was concluded that controlling for background variables was not a correct method to compute Canadian trends. The marginal trend seemed more accurate.

CZECH REPUBLIC

The conditional trend in science was less positive than the marginal trend.

Students in PISA 2003 were on average two months older than in PISA 2000 and their HISEI was two points higher (on a scale from 16 to 90). Removing these regressors (and the indicators for

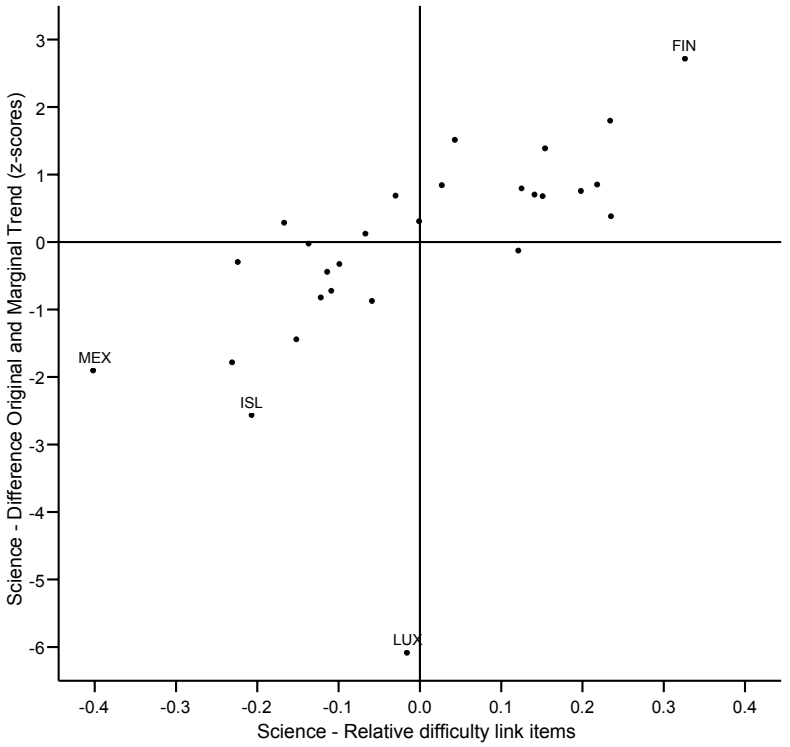

Figure 5. Scatterplot between relative difficulty of science link items versus difference in original and marginal trends in science

missing data) made the conditional trend equivalent to the marginal trend. Controlling for these changes in background variables probably gave a better estimate of the trend in science. Changes in these background variables had the same effect on the reading trend, but both marginal and conditional trends in reading were not significant.

GERMANY

The conditional trends were more positive than the marginal trends.

Ten percent of the students in PISA 2003 and three percent of the students in PISA 2000 had missing values for HISEI. The indicator for missing HISEI had a negative relationship with reading and science performance. Removing these students from the analysis resulted in an increase so that the conditional trends were approximately equivalent to the marginal trends. Since there was neither a good reason for removing these students nor for taking the effect of missing values into account, the marginal trends (including all students) seemed better trend estimates than the conditional estimates.

ICELAND

The conditional trends were more negative than the marginal trends.

Even though the significance levels were equal for the different trends, the size of the difference was not negligible (difference in z-scores was 1.72 for reading and 2.20 for science). No students had missing values for sex and age in 2003 (one percent and five percent in PISA 2000). Removing the students with missing values for these variables in PISA 2000 made the conditional trend approximately equal to the marginal trend. Since there was no good reason for deleting these students, the marginal trends were better indicators than the conditional trends.

JAPAN

The conditional trends were more negative than the marginal trends.

In PISA 2003, 63 percent of students had missing values for HISEI, while only 11 percent had missing values in PISA 2000. Since the missing indicator was negatively related to performance, the conditional trends were underestimations of the real trend.

NEW ZEALAND

The conditional trends were more negative than the marginal trends.

Students in PISA 2003 were on average one month older than students in PISA 2000. Removing age from the regression analysis to estimate the conditional trends resulted in less negative conditional trends, equivalent to the marginal trends. Therefore, the change in age caused the conditional trend to be different from the marginal trend and controlling for age seemed appropriate when computing trends in New Zealand.

There was also a change in the amount of missing values for HISEI and language at home (four and five percent in PISA 2000 and 14 and one percent in PISA 2003, respectively), but these effect cancelled each other out.

UNITED STATES OF AMERICA

The conditional trends were more negative than the marginal trends.

When compared to the PISA 2000 data the PISA 2003 data set had four percent more boys, six percent less missing data for sex, five percent less missing data for age, on average almost two months older students, three points higher HISEI (on a scale from 16 to 90) and nine percent less missing values for HISEI (see Table 6). Conditional trends seemed more appropriate for the USA than marginal trends because of these changes in distributions of background variables.

In summary, Table 7 highlights the best of the alternative trends where differences between marginal and conditional trends were judged as substantial.

Discussion

This study shows some interesting findings about trend estimation in international comparative research. Trends receive widespread attention from researchers, policy makers and the press. Within countries it is of interest to see if performance changes over time possibly due to changes in the educational system or due to methodological artefacts.

The analyses in this paper provide a careful and detailed examination of alternative ways of estimating trends and explaining differences between the results of these methods on a country-by-country level. The results show that a common approach to estimating trends for all countries may well be misleading.

First, it was found that a substantial amount of variation in difference between two sets of trends (original and marginal) could be accounted for by a form of country-by-item interaction. One of these sets of trends was based on international values for item parameters while the other used national item parameters. The relative average difficulty of the set of link items within a domain (compared to the average difficulty of the set of unique items) is not stable across countries. Using international item parameters leads for some countries to an underestimation and for others to an overestimation of the trend compared to their nationally estimated trends. Using national item parameters for estimating trends does not

Table 6

Descriptives of background variables for students from the USA in PISA 2000 and PISA 2003

Cycle	Boys (%)	Missing sex (%)	Age (months)	Missing age (%)	HISEI	Missing HISEI (%)	Language (%)	Missing language (%)
P2000	.46	6	188	5	52	15	10	6
P2003	.50	0	190	0	55	6	9	4

solve all problems because this will decrease the comparability between countries. Nevertheless, this form of country-by-item interaction can be accounted for when estimating trends using international parameters and is expected to improve the estimation of trends within countries.

Second, changes in distributions of background variables between surveys can have an effect on the estimation of trends if these variables are related to performance. This effect can either be a true change in the population or a reflection of something else, such as sampling variation or a change in the instrumentation. This leads to a difficulty in the estimation of trends because careful examination of the country's samples is necessary to conclude if an observed change in the distribution of a non-malleable variable should be adjusted for or not.

In conclusion, the national characteristics as described above could help improving the accuracy of estimating trends. The effect of item-by-country interaction on trends and testing a method for controlling for this effect is an important issue for future research. In addition to taking into

Table 7

As Table 4, but highlighting the better of the two alternative trends

	Reading			Science		
	Original	Marginal	Conditional	Original	Marginal	Conditional
Australia	0	0	0	0	0	- -
Austria	-	- - -	-	- - -	- - -	- - -
Belgium	0	0	0	+++	+++	+++
Canada	0	0	+++	0	0	++
Czech Republic	0	0	0	+++	++	0
Denmark	0	0	0	0	0	0
Finland	- -	0	0	+++	0	0
France	0	0	0	+++	+++	+++
Germany	+++	0	+++	+++	+++	+++
Hungary	0	0	0	0	0	0
Iceland	- - -	- - -	- - -	0	0	0
Ireland	-	0	0	0	0	0
Italy	- - -	- -	- - -	0	0	0
Japan	- -	0	- - -	- -	- -	- - -
Korea	+++	+++	+++	- -	- - -	- - -
Luxembourg	+++	+++	+++	+++	+++	+++
Mexico	- - -	- - -	- - -	- - -	- - -	- - -
Netherlands	- - -	- - -	- - -	0	0	0
New Zealand	0	- -	- - -	0	-	- - -
Norway	- -	- - -	- - -	- - -	- - -	- - -
Poland	+++	+++	+++	+	+++	+++
Portugal	0	+	0	++	++	+
Russian Federation	- - -	- - -	- - -	+++	+++	++
Spain	- - -	- - -	- - -	- - -	- -	- - -
Sweden	0	- - -	- - -	0	0	0
Switzerland	++	+	0	+++	+++	+++
United Kingdom	- - -	- - -	- - -	- -	- - -	- - -
United States	0	0	- - -	0	0	- -

Note: Significance level: 2003 better than 2000: 2003 worse than 2000:

$p > .10$	0	0
$p < .10$	+	-
$p < .05$	++	- -
$p < .01$	+++	- - -

account the effect of unwanted changes in background variables, PISA soon will have collected data at more than two points in time, which will also smooth out uncertainties in trends that are caused by sampling and other issues.

References

Adams, R. J., Wu, M. L., and Macaskill, G. (1998). Scaling methodology and procedures for the mathematics and science scales. In M. O. Martin and D. Kelly (Eds.). *TIMSS technical report volume II: Implementation and analysis (primary and middle school years)* (pp. 147-174). Boston: Boston College.

Adams, R. J. (2002). Scaling PISA cognitive data. In R. J. Adams and M. Wu (Eds.), *Programme for international student assessment: PISA 2000 technical report* (pp. 99-108). Paris: OECD Publications.

Adams, R. J., and Carstensen, C. (2002). Scaling outcomes. In R. J. Adams and M. Wu (Eds.), *Programme for international student assessment: PISA 2000 technical report* (pp. 149-162). Paris: OECD Publications.

Adams, R. J. and Wu, M. L. (2002). (Eds.), *Programme for international student assessment: PISA 2000 technical report* (pp. 99-108). Paris: OECD Publications.

Adams, R. J., Wu, M. L., and Carstensen, C. H. (2006). Application of multivariate Rasch models in international large scale educational assessment. In M. v. Davier and C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 271-280). New York: Springer Verlag.

Ganzeboom, H. B. G., de Graaf, P. M., and Treiman, D. J. (1992). A standard international socioeconomic index of occupational status. *Social Science Research*, *21*, 1-56.

Hedges, L. V., and Vevea, J. L. (2003). *NAEP validity studies: A study of equating in NAEP (NCES 2003-13).* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Johnson, E. G., and Carlson, J. E. (1992). *NAEP technical report*. Washington, DC: National Center for Educational Statistics.

Judkins, D. R. (1990). Fay's method of variance estimation. *Journal of Official Statistics*, *3,* 223-239.

Kim, S. H., and Cohen, A. S., (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*(2), 131-143

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago: MESA Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Michaelides, M. P., and Haertel, E. H. (2004). *Sampling of common items: An unrecognized source of error in test equating* (CSE Report 636). Los Angeles: The Regents of the University of California.

Mislevy, R. J., and Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *The NAEP 1983-1984 technical report.* (pp. 293-360). Princeton, NJ: Educational Testing Service.

Monseur, C., and Berezner A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement*, *8*, 323-335.

Mullis, I. V. S., and Martin, M. (1998). Item analysis and review. In M. O. Martin and D. Kelly (Eds.). *TIMSS technical report volume II: Implementation and analysis (Primary and middle school years)* (pp. 111-146). Boston: Boston College.

OECD (2001). *Knowledge and skills for life: First results from PISA 2000.* Paris: OECD Publications.

OECD (2004). *Learning for tomorrow's world: First results from PISA 2003.* Paris: OECD Publications.

OECD (2005a). *PISA 2003 data analysis manual—SPSS users*. Paris: OECD Publications.

OECD (2005b). *Programme for international student assessment: PISA 2003 technical report.* Paris: OECD Publications.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press.)

Rubin, D. B. (1987). *Multiple imputations for non-response in surveys.* New York: Wiley.

Wu, M. L., Adams, R. J., and Wilson, M. R. (1998). ACER ConQuest: Generalized item response modeling software [Computer program]. Melbourne: ACER Press.

Appendix A—Reading

	ORIGINAL TREND	MARGINAL TREND			CONDITIONAL TREND		
Country	*Z*–score	Unstandardized regression coefficient	*S.E.*	*Z*–score	Unstandardized regression coefficient	*S.E.*	*Z*–score
Australia	–0.554	–0.003	0.050	–0.064	–0.011	0.039	–0.284
Austria	–1.948	–0.142	0.049	–2.924	–0.071	0.040	–1.781
Belgium	0.481	0.063	0.055	1.154	0.026	0.048	0.546
Canada	–1.339	0.043	0.027	1.596	0.077	0.023	3.369
Czech Republic	–0.505	0.043	0.053	0.805	–0.028	0.036	–0.786
Denmark	1.174	–0.017	0.040	–0.423	–0.051	0.036	–1.413
Finland	–2.049	0.014	0.039	0.354	0.005	0.037	0.136
France	–1.465	0.035	0.046	0.760	–0.001	0.036	–0.028
Germany	2.593	0.033	0.050	0.664	0.122	0.044	2.813
Hungary	0.045	0.035	0.055	0.642	0.039	0.115	0.342
Iceland	–6.572	–0.108	0.027	–4.080	–0.155	0.027	–5.796
Ireland	–1.647	0.036	0.048	0.752	0.035	0.041	0.874
Italy	–3.933	–0.118	0.048	–2.426	–0.126	0.045	–2.831
Japan	–2.338	–0.106	0.073	–1.466	–0.279	0.073	–3.826
Korea	3.842	0.166	0.044	3.767	0.114	0.038	3.013
Luxembourg	18.553	0.381	0.026	14.545	0.300	0.025	11.789
Mexico	–5.154	–0.338	0.062	–5.452	–0.292	0.051	–5.690
Netherlands	–3.546	–0.193	0.051	–3.809	–0.157	0.040	–3.896
New Zealand	–1.009	–0.086	0.044	–1.966	–0.117	0.041	–2.861
Norway	–2.155	–0.203	0.048	–4.222	–0.244	0.044	–5.547
Poland	2.982	0.172	0.062	2.751	0.144	0.055	2.638
Portugal	0.520	0.122	0.071	1.723	0.082	0.059	1.390
Russia	–3.984	–0.174	0.067	–2.611	–0.203	0.059	–3.413
Spain	–3.900	–0.118	0.045	–2.619	–0.137	0.036	–3.796
Sweden	–0.261	–0.177	0.038	–4.589	–0.164	0.031	–5.308
Switzerland	1.980	0.115	0.063	1.812	0.085	0.054	1.570
United Kingdom	–3.506	–0.245	0.039	–6.346	–0.214	0.032	–6.594
United States	–1.189	–0.088	0.096	–0.924	–0.212	0.067	–3.173

Appendix B—Science

	ORIGINAL TREND	MARGINAL TREND			CONDITIONAL TREND		
Country	*Z*–score	Unstandardized regression coefficient	*S.E.*	*Z*–score	Unstandardized regression coefficient	*S.E.*	*Z*–score
Australia	0.414	–0.061	0.044	–1.384	–0.078	0.035	–2.234
Austria	–5.457	–0.239	0.047	–5.131	–0.172	0.038	–4.570
Belgium	3.564	0.185	0.050	3.690	0.152	0.044	3.421
Canada	–0.879	0.015	0.027	0.562	0.053	0.022	2.359
Czech Republic	3.127	0.108	0.044	2.439	0.020	0.037	0.558
Denmark	1.459	0.003	0.040	0.071	–0.031	0.033	–0.927
Finland	4.014	0.040	0.031	1.295	0.028	0.029	0.967
France	3.856	0.149	0.047	3.154	0.119	0.040	2.963
Germany	4.202	0.172	0.049	3.522	0.243	0.042	5.842
Hungary	0.399	0.059	0.053	1.121	0.012	0.109	0.106
Iceland	–1.059	0.034	0.023	1.503	–0.017	0.024	–0.695
Ireland	–0.556	–0.038	0.044	–0.863	–0.038	0.036	–1.074
Italy	0.341	–0.019	0.046	–0.416	–0.026	0.043	–0.604
Japan	–2.025	–0.157	0.079	–2.001	–0.294	0.083	–3.551
Korea	–2.409	–0.135	0.050	–2.695	–0.223	0.048	–4.635
Luxembourg	14.887	0.460	0.022	20.970	0.388	0.023	16.953
Mexico	–5.735	–0.199	0.052	–3.833	–0.160	0.043	–3.754
Netherlands	–0.094	–0.084	0.052	–1.608	–0.062	0.044	–1.396
New Zealand	–0.883	–0.067	0.039	–1.726	–0.096	0.037	–2.599
Norway	–4.579	–0.120	0.043	–2.797	–0.157	0.039	–3.989
Poland	1.823	0.161	0.061	2.645	0.144	0.054	2.676
Portugal	2.505	0.140	0.059	2.380	0.096	0.051	1.897
Russia	3.588	0.199	0.073	2.735	0.168	0.067	2.512
Spain	–2.924	–0.110	0.044	–2.483	–0.126	0.035	–3.607
Sweden	–0.475	0.015	0.038	0.397	0.027	0.032	0.829
Switzerland	3.201	0.183	0.065	2.818	0.148	0.056	2.657
United Kingdom	–2.462	–0.123	0.038	–3.257	–0.113	0.031	–3.696
United States	–0.751	–0.041	0.089	–0.456	–0.158	0.063	–2.507