

## **Use of Different Models for Estimating Trends**

*Eveline Gebhardt (Australian Council for Educational Research) and  
Raymond J. Adams (University of Melbourne, Australia and Australian Council for  
Educational Research )*

Paper presented at the Annual Meeting of The American Educational Research  
Association. San Francisco, CA, April 7-11, 2006

## Abstract

The Programme or International Student Achievement (PISA) has now collected two waves of data, the first in 2000 and the second in 2003. There are 33 countries that participated in both studies and for which trend data in reading literacy performance is available. In this paper we explore the extent to which the outcomes of trend analyses are sensitive to the choice of test equating methodologies, the choice of regression models and the choice of linking items.

Deleted: and

To establish trends PISA equated its 2000 and 2003 tests using a methodology that involved estimating linear transformations that mapped 2003 item response theory (IRT)-scaled scores to the previously established PISA 2000 IRT-scaled scores. In this paper we compare the outcomes of this approach with an alternative, which involves the joint IRT 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.

Deleted: IRT

Deleted: .

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 sub-group level.

## Introduction

The Organisation for Economic Cooperation and Development's Programme for International Student Assessment (OECD PISA) is a collaborative effort, involving all OECD countries and a significant number of partner countries, to measure how well 15-year-old students are prepared to meet the challenges of today's knowledge societies. The assessment looks to the future, focusing on young people's ability to use their knowledge and skills to meet real-life challenges rather than on the mastery of specific school curricula. The term *literacy* is used to encapsulate this broader concept of knowledge and skills. The age of 15 is used because in most OECD countries students are approaching the end of compulsory schooling (OECD, 2005a).

Deleted: OECD's

PISA is an ongoing survey with a data collection every three years. Each three-year period is referred to as a cycle. The first PISA survey was conducted in 2000 in 32 countries, using written tasks answered in schools under independently supervised test conditions following consistently applied standards. Another 11 countries participated in the same survey in late 2001 or early 2002. The second survey was conducted in 2003 in 41 countries. Table 1 gives the list of participating countries for PISA 2000 and PISA 2003.

	PISA 2000	PISA 2003
OECD countries	Australia, Austria, Belgium, Canada, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Korea, Luxembourg, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Spain, Sweden, Switzerland, United Kingdom, United States	Australia, Austria, Belgium, Canada, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Korea, Luxembourg, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Spain, Sweden, Switzerland, Turkey, United Kingdom, United States
Partner countries	Albania, Argentina, Brazil, Bulgaria, Chile, Hong-Kong-China, Indonesia, Israel, Latvia, Liechtenstein, Macedonia, Peru, Romania, the Russian Federation, Thailand	Brazil, Hong-Kong-China, Indonesia, Liechtenstein, Latvia, Macao-China, the Russian Federation, Thailand, Tunisia, Uruguay, Serbia

**Table 1: Countries participating in PISA 2000 and PISA 2003**

PISA assesses reading, mathematical, and scientific literacy (referred to as reading, mathematics and science in the remainder of this paper). For each data collection, one of these three domains is chosen as the major domain, while the others are considered as minor domains. PISA 2000 focused on reading, while the major domain for PISA 2003 was mathematics. Approximately half of the testing time is devoted to the major domain and the minor domains share the remainder.

PISA 2000 used nine booklets and PISA 2003 used 13 booklets and in each case the booklets contained approximately 50 items. In both cycles one special booklet was constructed for students with special educational needs. Using this booklet for assessment in special schools was optional and not used by all countries. Booklet allocation was rotated systematically in classes. However, variation in mean performance by booklet was sometimes larger than expected. Therefore, booklet corrections were applied following procedures described in Adams and Carstensen (2002) and OECD (2005b, pp. 198-206).

**Deleted:** with each approximately 50 items

To scale the PISA data, the mixed coefficients multinomial logit model, as described by Adams, Wilson and Wang (1997), was used and implemented by ConQuest software (Wu, Adams and Wilson, 1997). The model is a generalised form of the Rasch model. Dichotomous items were scaled with Rasch's simple logistic model (Rasch, 1960), and items with multiple score categories were scaled with Masters's partial credit model (Masters, 1982).

In the field trials of both cycles, several item characteristics were examined before a final selection was made for assessment in the main study. Among these characteristics were item-by-country interactions. Item parameters were estimated for each country and variation in national parameters of one item across countries was used to judge the international comparability of that item.

After judging item-by-country interactions and selecting the final set of items for the main studies, final IRT scaling was implemented in two steps: (1) international item calibration and (2) national student score generation using all student background information collected with questionnaires as conditioning variables. The first step used an international calibration sample with 500 randomly selected students from each OECD country and their responses to the test items. Marginal maximum likelihood estimation was used, with the assumption that students were sampled from a multivariate normal distribution. In the second step, student background variables were used to generate student scores country by country, while the item parameters were anchored to the international parameters from step one.

In this second step, plausible value methodology was applied for estimating student abilities, because the primary concern was with the estimation of population parameters and not individual performances (see Adams, Wu, & Carstensen, to appear, for a more detailed description). This approach was developed by Mislevy and Sheehan (1987, 1989) and based on the imputation theory of Rubin (1987). For each student a posterior distribution was estimated using the student's raw score on a domain within a booklet, information about the student's background collected with questionnaires, the school's adjusted mean performance on the major domain and student's scores on the other assessed domains. Five random plausible values for each student were drawn from its individual posterior distribution. Subsequently, these plausible values were used for further analysis.

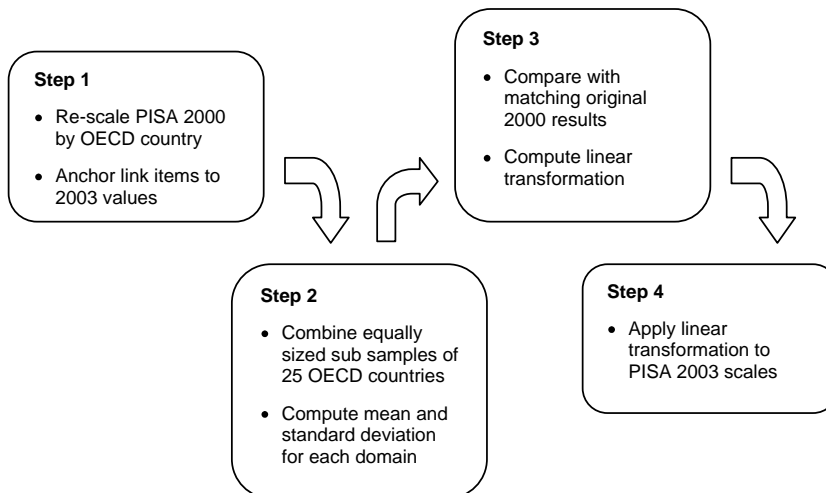
Not all booklets contained items from all domains. That is, individual students do not always respond to items from all domains. For each domain, the published results of PISA 2000 were estimated using only students who responded to items in a domain, while PISA 2003 used all sampled students (OECD, 2005b). The inclusion of all students in the analyses reduces the sampling error in the estimation of population parameters. This occurs because the dimensions are typically very highly correlated. In the new analyses for this paper all students were used regardless of the booklet they were assigned.

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.

The steps involved in the original linking of PISA 2000 and PISA 2003 reading and science scales were as follows (see also Figure 1).

- Step 1. The PISA 2000 data from each of the OECD countries were re-scaled with full conditioning and with link items anchored at their PISA 2003 values.
- Step 2. The mean and standard deviation of each domain were calculated for a combined data set of 25 equally weighted OECD countries.
- 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. 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. Linear transformation from step 3 was applied to the PISA 2003 scales.



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

There is some debate about including an often overlooked source of error caused by the sample of link items (U.S. Department of Education, National Center for Education Statistics, 2003 and Michaelides & Haertel, 2004). The transformation that equates the PISA 2000 and PISA 2003 data depends upon the change in difficulty of each of the individual link items and consequently the sample of link items that was chosen will influence the choice of transformation. This means that if an alternative set of link items had been chosen the resulting transformation would be slightly different. The consequence is an uncertainty in the transformation due to the sampling of the link items, just as there is an uncertainty in values such as country means due to the use of a sample of students. However, some researchers disagree with including this form of error, because items are fixed effects, therefore, no item sampling is assumed.

The uncertainty that results from the link-item sampling is referred to as linking error. Just as with the error that is introduced through the process of sampling students, the exact magnitude of this linking error cannot be determined. However, the likely range of magnitudes for this error could be estimated and was taken into account when interpreting and reporting original PISA results (OECD, 2005b, pp. 133-134 and 211-214). As with sampling errors, the likely range of magnitude for the errors is represented as a standard error. However, this equating error is not used in the current study for mainly practical reasons.

Trends are important information for participating countries. They change ranking of countries over time and they indicate if countries performances have changed within countries. However, several methodologies can be applied to equate PISA 2003 scores to the PISA 2000 scale, each with their own advantages and disadvantages and each with at least slightly different results.

Three issues in trend methodology are examined in this paper. (a) National item parameters can be used instead of international item parameters to create a more accurate picture of trends within countries. Consequently, each country will have

different item parameters, which will make comparisons across countries less acceptable. (b) Data sets of the two cycles can be merged so that items are calibrated on a joint data set and no linear transformations are needed for equating. (c) In the estimation of trends, changes in the sample or population could be accounted for. For example, when for some reason two data sets from different cycles have different proportions of boys, it is unlikely, although not impossible, that this reflects a real change in the population. Therefore, controlling for the proportion of boys in the calculation of trends could lead to a more accurate measure of the trend in a particular country. On the contrary, imagine a country with a positive trend, not adjusted for possible changes in background characteristics of the students, a proportion of boys that is stable across cycles, but an increase in missing values for sex. In case missing values for sex are negatively related to performance, 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.

Unadjusted trends are called *marginal* and adjusted trends *conditional* in this paper. In case the marginal and conditional trends differ from each other, the reasons for this difference are being explored in a particular country and a decision is made about the accuracy of the two trends. .

Two alternative trend estimates were computed and compared to the original results and to each other (see Table 2). 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.

Item parameters	Equating method	Trend Indicator	
		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	—	—

Deleted: o

**Table 2: Variables in trend methodology**

## Methodology

Scaling for this paper was done 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 and 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, item parameters were re-estimated in uni-dimensional models, using the combined data set 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. 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 model was run with mathematics, reading and science as domains to draw plausible values anchoring items to the parameters that were derived in the first step. All students (including special education) and all link and non-link items were included in the model. Deviation booklet contrast coding and a dummy for cycle were used as regressors. These had to be used to make the three-dimensional model mathematically equivalent to the uni-dimensional models with booklets as facets.<sup>1</sup> Therefore, they 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 student 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 regressors as well. These common background variables were sex, highest educational status of parents (HISEI), age, language at home, school mean in mathematics, reading and science, mother's occupation, and father's occupation and their indicators for missing values.

Deleted: .

Finally, single and multiple regression analyses were run in SPSS (using student sampling weights and BRR replication method) to compute 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).

---

<sup>1</sup> Note that the ConQuest cannot be easily configured to permit different facet effects for each dimension.

Not all common variables that were included in the conditioning model to draw plausible values were included in this multiple regression analysis. School means were not included because change in school means can reflect a true trend.

Implementation of an educational program~~me~~ in part of a country's schools may increase their performance. We do not wish to control for these changes in performance. Parents' occupational status is excluded as well, because it is highly correlated with parents' highest educational status (HISEI). The (partial) regression coefficient of the cycle dummy was used as the conditional trend for a domain within a country.

Deleted: '

As mentioned in the introduction, estimates of linking errors were not included in the standard errors. One reason is that no consensus is 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 3.

**Table 3: 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
AUS	5176	12551	229152	235591	528	(3.5)	526	(2.2)	528	(3.5)	529	(2.1)
AUT	4745	4597	71547	85931	507	(2.4)	498	(3.8)	519	(2.5)	494	(3.6)
BEL	6670	8796	110095	111831	507	(3.6)	509	(2.6)	496	(4.3)	513	(2.4)
CAN	29687	27953	348481	330436	534	(1.6)	531	(1.8)	529	(1.6)	527	(2.0)
CHE	6100	8420	72010	86491	494	(4.2)	505	(3.1)	496	(4.4)	514	(3.6)
CZE	5365	6320	125639	121183	492	(2.4)	489	(3.6)	511	(2.4)	524	(3.3)
DEU	5073	4660	826816	884358	484	(2.5)	495	(3.3)	487	(2.4)	505	(3.6)
DNK	4235	4218	47786	51741	497	(2.4)	501	(2.8)	481	(2.8)	487	(3.0)
ESP	6214	10791	399055	344372	493	(2.7)	478	(2.5)	491	(3.0)	479	(2.6)
FIN	4864	5796	62826	57884	546	(2.6)	540	(1.7)	538	(2.5)	550	(1.9)
FRA	4673	4300	730494	734579	505	(2.7)	499	(2.6)	500	(3.2)	518	(3.1)
GBR	9340	9535	643041	698579	523	(2.6)	511	(2.4)	532	(2.7)	523	(2.7)
HUN	4887	4765	107460	107044	480	(4.0)	480	(2.5)	496	(4.2)	498	(2.7)
IRL	3854	3880	56209	54850	527	(3.2)	520	(2.5)	513	(3.2)	511	(2.6)
ISL	3372	3350	3869	3928	507	(1.5)	494	(1.4)	496	(2.2)	493	(1.5)
ITA	4984	11639	510792	481521	487	(2.9)	471	(3.0)	478	(3.1)	479	(3.2)
JPN	5256	4707	1000000	1000000	522	(5.2)	507	(3.7)	550	(5.5)	536	(4.2)
KOR	4982	5444	579109	533504	525	(2.4)	540	(3.1)	552	(2.7)	541	(3.5)
LUX	3404	3923	3981	4080	441	(1.6)	479	(1.2)	443	(2.3)	483	(1.4)
MEX	4600	29983	960011	1000000	422	(3.3)	394	(4.2)	422	(3.2)	394	(3.7)
NLD	2503	3992	157327	184943	532	(3.4)	516	(2.9)	529	(4.0)	529	(3.2)
NOR	4147	4064	49579	52816	505	(2.8)	497	(2.7)	500	(2.7)	482	(3.0)
NZL	3667	4511	46757	48638	529	(2.8)	525	(2.7)	528	(2.4)	525	(2.5)
POL	3654	4383	542005	534900	479	(4.5)	495	(2.9)	483	(5.1)	494	(3.0)
PRT	4585	4608	99998	96857	470	(4.5)	473	(3.7)	459	(4.0)	472	(3.4)
RUS	6701	5974	2000000	2000000	462	(4.2)	439	(3.9)	460	(4.7)	483	(4.2)
SWE	4416	4624	94338	107104	516	(2.2)	515	(2.5)	512	(2.5)	510	(2.8)
USA	3846	5456	3000000	3000000	504	(7.0)	495	(3.0)	499	(7.3)	494	(3.0)

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 due to some observed inconsistencies in students' weights. The Netherlands and United Kingdom were not included in the official PISA reports, because they had non-response difficulties in one of the cycles. They were 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 1 presents the significance of the differences for each country between PISA 2000 and PISA 2003 in reading and science (see [Appendix](#) for regression coefficients,

**Deleted:** Two OECD countries that participated in both cycles were not included.

**Deleted:** had

**Deleted:** because of inaccurate population counts in PISA 2000. Liechtenstein is not included because of its small sample size.

**Deleted:** All other OECD countries were included as well as some of the partner countries.

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.

**Table 4: A comparison of the significance of trends between three alternative methods for equating**

Deleted: 3

	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:**

<b>p &gt; .10</b>	<b>0</b>	<b>0</b>
<b>p &lt; .10</b>	+	-
<b>p &lt; .05</b>	++	--
<b>p &lt; .01</b>	+++	---

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

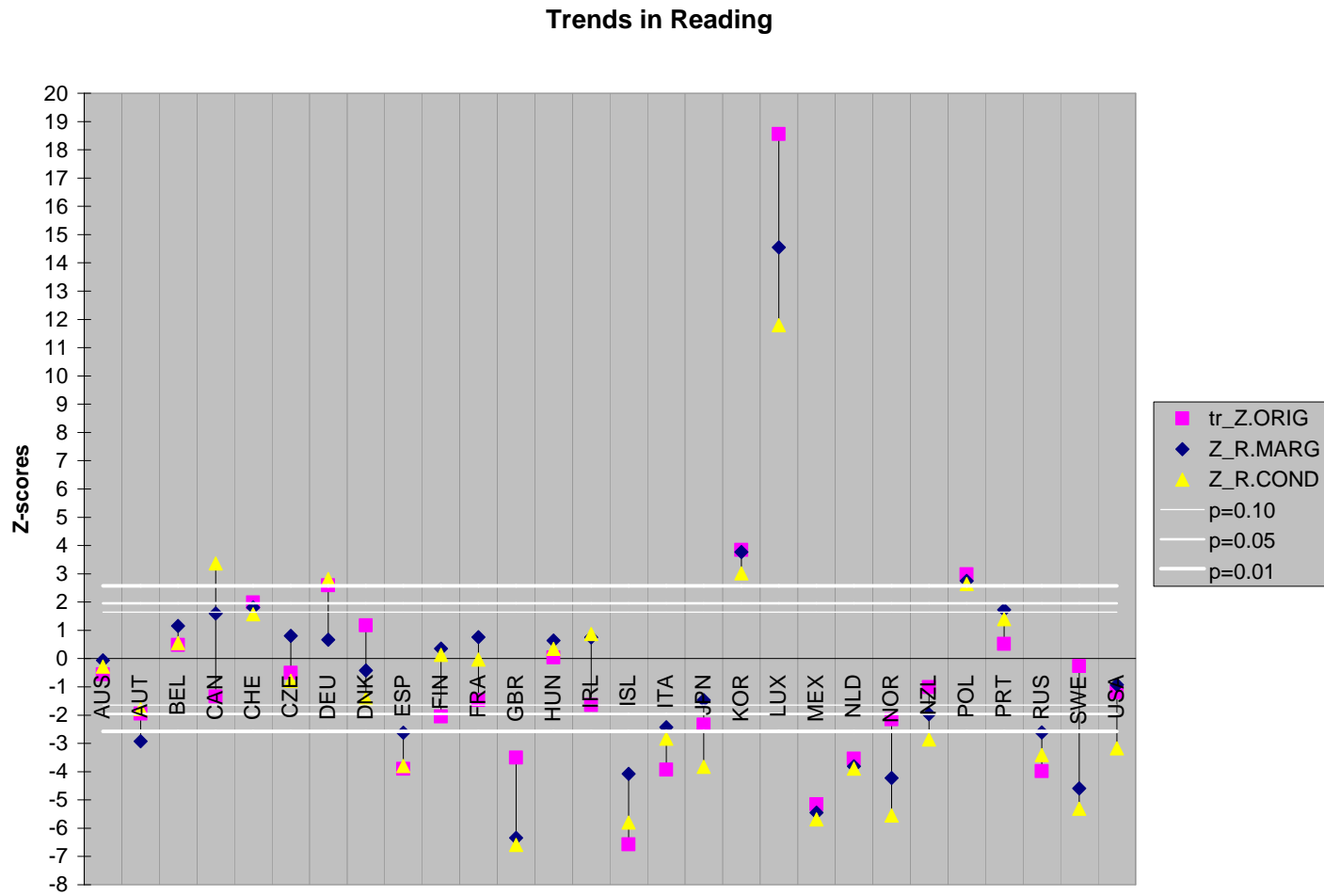
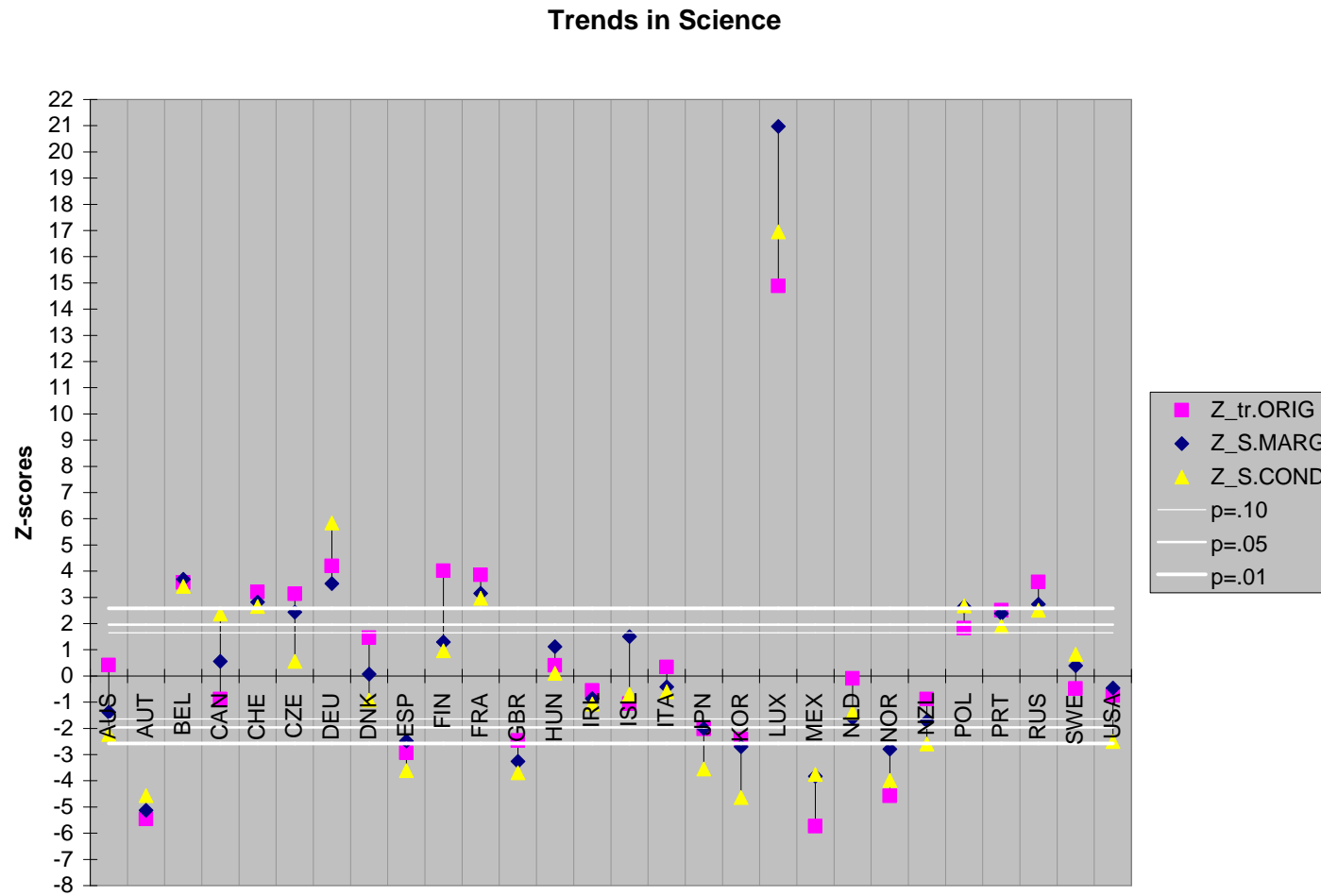


Figure 3: Three alternative trends between PISA 2000 and PISA 2003 in science by country



### 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 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 ( $p < .01$ ). 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 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 item-by-country interaction. These average reading item difficulties and the relative difficulty of the set of link items are displayed in Table 5, followed by an explanation of the figures in the table.

Deleted: .

**Table 5: Computation of relative difficulty of reading link items**

	Reading				RELATIVE DIFFICULTY LINKS
	Unique 2000	Links	National relative difficulty links	(Adjusted) international relative difficulty	
International 2000	0.01	-0.03		0.04	
International 2003		-0.41			
AUS	0.01	-0.04	0.05	0.04	0.01
AUT	0.02	-0.06	0.07	0.04	0.03
BEL	-0.01	0.03	-0.04	0.04	-0.08
CAN	-0.02	0.05	-0.07	0.04	-0.11
CHE	0.01	-0.03	0.04	0.04	0.00
CZE	-0.01	0.04	-0.05	0.04	-0.09
DEU	0.01	-0.05	0.06	0.04	0.02
DNK	0.03	-0.12	0.15	0.04	0.11
ESP	-0.02	0.08	-0.10	0.05	-0.15
FIN	0.01	-0.04	0.05	0.04	0.01
FRA	-0.02	0.07	-0.09	0.04	-0.13
GBR	0.01	-0.12	0.14	0.04	0.10
HUN	-0.03	0.09	-0.11	0.03	-0.14
IRL	-0.01	0.02	-0.03	0.04	-0.07
ISL	-0.01	0.00	-0.01	0.03	-0.04
ITA	-0.03	0.09	-0.12	0.02	-0.14
JPN	-0.03	0.09	-0.12	0.04	-0.16
KOR	0.01	-0.04	0.06	0.08	-0.03
LUX	-0.01	0.03	-0.04	0.04	-0.08
MEX	-0.05	0.18	-0.23	0.04	-0.27
NLD	-0.01	-0.04	0.03	0.03	0.00
NOR	0.02	-0.09	0.11	0.04	0.08
NZL	0.03	-0.10	0.13	0.04	0.09
POL	0.00	0.07	-0.07	0.01	-0.08
PRT	-0.04	0.14	-0.18	0.04	-0.22
RUS	-0.04	0.11	-0.15	0.05	-0.21
SWE	0.06	-0.20	0.26	0.05	0.21
USA	0.02	0.01	0.01	0.04	-0.03

The second column in Table 5 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 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. 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 shaded columns in Table 5 and Table 6) 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 6 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 for items that were nationally deleted. 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 6.

Table 6: Computation of relative difficulty of science link items

	Science							RELATIVE DIFFICULTY LINKS
	Unique 2000	Unique 2003	Links	Relative difficulty links 2000	Relative difficulty links 2003	National relative difficulty links	(Adjusted) international relative difficulty	
International 2000	0.08		-0.03	0.10				
International 2003		0.08	0.10		-0.02		0.13	
AUS	0.23	-0.13	-0.04	0.27	-0.10	0.36	0.13	0.23
AUT	-0.03	-0.06	0.03	-0.07	-0.09	0.03	0.13	-0.10
BEL	0.11	-0.13	0.01	0.11	-0.14	0.25	0.13	0.12
CAN	0.04	0.07	-0.04	0.08	0.11	-0.03	0.13	-0.15
CHE	0.17	-0.19	0.01	0.17	-0.20	0.36	0.13	0.24
CZE	0.09	-0.01	-0.03	0.12	0.02	0.10	0.13	-0.03
DEU	0.11	-0.25	0.05	0.08	-0.19	0.27	0.12	0.15
DNK	0.12	-0.16	0.01	0.11	-0.17	0.28	0.13	0.15
ESP	-0.01	-0.02	0.01	-0.02	-0.03	0.01	0.13	-0.11
FIN	0.13	-0.33	0.07	0.05	-0.40	0.45	0.13	0.33
FRA	0.16	-0.11	-0.02	0.18	-0.09	0.27	0.13	0.14
GBR	0.17	-0.08	-0.03	0.21	-0.05	0.25	0.13	0.12
HUN	-0.03	-0.05	0.03	-0.06	-0.08	0.02	0.13	-0.11
IRL	0.06	-0.07	0.00	0.05	-0.07	0.13	0.13	0.00
ISL	-0.05	-0.06	0.04	-0.03	0.13	-0.16	0.05	-0.21
ITA	0.17	-0.15	-0.01	0.18	-0.14	0.32	0.13	0.20
JPN	0.02	0.03	-0.02	0.04	0.05	-0.01	0.13	-0.14
KOR	-0.04	0.00	0.01	-0.05	-0.01	-0.04	0.13	-0.17
LUX	0.06	-0.05	0.00	0.06	-0.05	0.11	0.13	-0.02
MEX	-0.03	0.25	-0.08	0.05	0.32	-0.28	0.13	-0.40
NLD	0.10	-0.13	0.01	0.11	-0.06	0.17	0.13	0.04
NOR	-0.09	0.01	-0.02	-0.07	0.04	-0.11	0.12	-0.23
NZL	0.08	-0.07	0.00	0.08	-0.07	0.15	0.13	0.03
POL	0.05	0.05	-0.04	0.09	0.09	0.00	0.13	-0.12
PRT	0.07	0.05	-0.09	0.16	0.14	0.01	0.08	-0.07
RUS	0.29	-0.09	-0.06	0.30	-0.03	0.33	0.11	0.22
SWE	0.05	-0.02	-0.01	0.06	-0.01	0.07	0.13	-0.06
USA	0.07	0.17	-0.09	0.16	0.25	-0.10	0.13	-0.22

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.

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

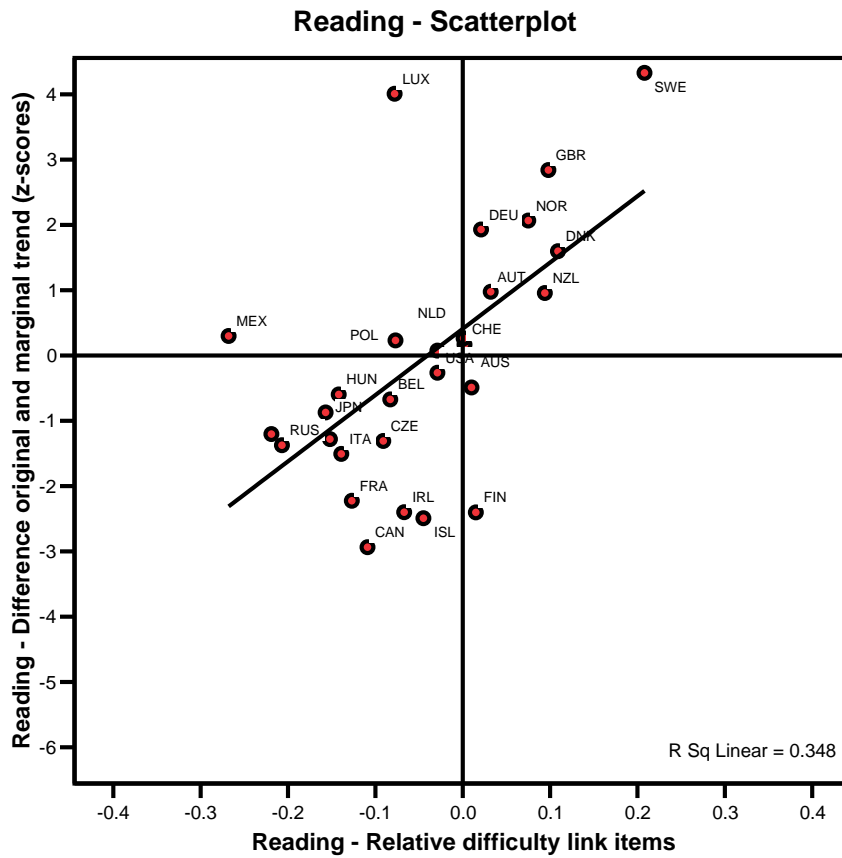
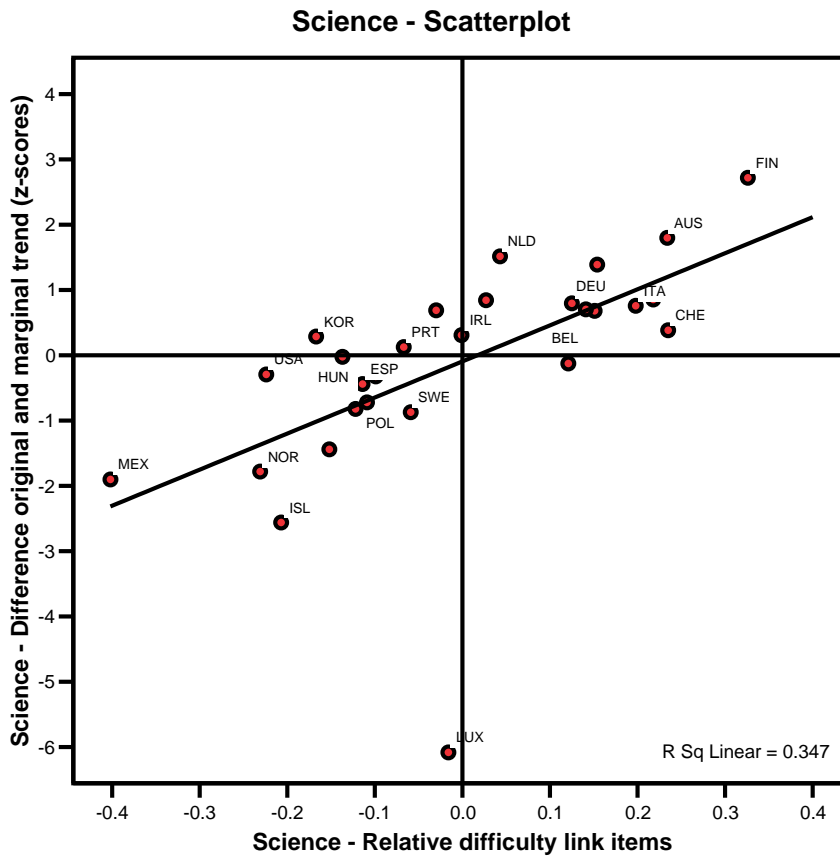


Figure 5: Scatterplot between relative difficulty of science link items versus difference in original and marginal trends in science



The correlation for reading was 0.59 and for science 0.58. Luxemburg is an outlier in both plots with an extreme positive trend. Deleting Luxemburg results in correlations of 0.67 and 0.82 respectively.

**Marginal versus Conditional trends**

It was assumed that changes in background variables were more likely to be changes in the sample than in the real population, because the two cycles were only three years apart. Therefore, when changes in background variables (other than missing indicators) were observed, the conditional trend was preferred above the marginal trend.

AUSTRALIA (AUS)

Although the trends in science show different levels of significance in [Table 4](#), Figure 3 shows that the actual difference was rather small. The change in Z-score is mainly caused by a change in standard error, not in regression coefficient (marginal regression coefficient is -0.06 (SE=0.035), conditional regression coefficient is -0.08 (SE=0.044)). The small difference between the marginal and conditional trends is mainly cause by an increase in age of one month (mean age is 188 months in PISA

Formatted: Default Paragraph Font  
Deleted: Table 4

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.*

Austria has 47 percent boys, a mean HISEI of 50 and six percent of students that do not speak the test language at home most of the time in PISA 2000. 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.

#### CANADA (CAN)

*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, while 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 (CZE)

*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 missing) 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.

Deleted:

#### GERMANY (DEU)

*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

effect on 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 (ISL)

*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.

#### ITALY (ITA)

Although the trends in reading show different levels of significance in [Table 4](#), Figure 2 shows that the actual difference was rather small.

Formatted: Default Paragraph Font

Deleted: Table 4

#### JAPAN (JPN)

*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 (NZL)

*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.

#### PORTUGAL (PRT)

Although the trends in science and reading show different levels of significance in [Table 4](#), Figure 2 and Figure 3 show that the actual differences were rather small.

Deleted: Table 4

Formatted: Default Paragraph Font

RUSSIAN FEDERATION (RUS)

Although the trends in science show different levels of significance in [Table 4](#), Figure 3 shows that the actual difference was rather small.

Deleted: Table 4  
Formatted: Default Paragraph Font

SPAIN (ESP)

Although the trends in science and reading show different levels of significance in [Table 4](#), Figure 2 and Figure 3 show that the actual differences were rather small.

Deleted: Table 4  
Formatted: Default Paragraph Font

SWITZERLAND (CHE)

Although the trends in science show different levels of significance in [Table 4](#), Figure 3 shows that the actual difference was rather small.

Formatted: Default Paragraph Font  
Deleted: Table 4

UNITED STATES OF AMERICA (USA)

*The conditional trends were more negative than the marginal trends.*

As shown in Table 7, the PISA 2003 data set had four percent more boys, six percent less missing values for sex and five 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 than the PISA 2000 data set. Conditional trends seemed more appropriate for the USA than marginal trends because of these changes in distributions of background variables.

**Table 7: 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

In summary, Table 8 highlights the best of the alternative trends where differences between marginal and conditional trends were judged as substantial.



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 is not a panacea, 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 trend if these variables are related to performance. This effect can either be a true change in the population or a reflection of something else. This is a difficulty in the estimation of trends, because careful examination of the country's samples is necessary conclude if it is a real population change or not. In case there is a real change in the population, controlling for changes in background variables will lead to an over- or underestimation of the real trend. However, if the change is a reflection of something else, they should be accounted for to improve the estimation of trends.

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 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. (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.
- 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.
- Adams, R. J., Wilson, M. R., and Wang, W. (1997). The multidimensional random coefficients multinomial logit model. Applied Psychological Measurement, 21, 1-24.
- Adams, Wu, & Carstensen, to appear. Chapter 17.
- Judkins, D. R. (1990). Fay's method of variance estimation. Journal of Official Statistics, 3, 223-239.
- Linacre, J. M. (1994). Many-Facet Rasch Measurement. Chicago: MESA.

- Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.
- Michaelides, M. P. & 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. & Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), The NAEP 1983-1984 Technical Report. Princeton: Educational Testing Service.
- Mislevy, R. J. & Sheehan, K. M. (1989). Information matrices in latent-variable models. Journal of Educational Statistics, 14, 335-350.
- OECD (2001). Knowledge and Skills for Life - First Results from PISA 2000. Paris: OECD.
- OECD (2005a). PISA 2003 Data Analysis Manual - SPSS users. Paris: OECD.
- OECD (2005b). Programme for International Student Assessment - PISA 2003 Technical Report. Paris: OECD.
- Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Nielsen and Lydiche.
- Rubin, D. B. (1987). Multiple Imputations for Non-Response in Surveys. New York: Wiley.
- U.S. Department of Education, National Center for Education Statistics (2003). NAEP Validity Studies: A Study of Equating in NAEP, NCES 2003-13, by L. V. Hedges and J. L. Vevea. Washington DC

## Appendix

READING	ORIGINAL TREND	MARGINAL TREND			CONDITIONAL TREND		
	Z-score	Unstandardized regression coefficient	S.E.	Z-score	Unstandardized regression coefficient	S.E.	Z-score
AUS	-0.554	-0.003	0.050	-0.064	-0.011	0.039	-0.284
AUT	-1.948	-0.142	0.049	-2.924	-0.071	0.040	-1.781
BEL	0.481	0.063	0.055	1.154	0.026	0.048	0.546
CAN	-1.339	0.043	0.027	1.596	0.077	0.023	3.369
CHE	1.980	0.115	0.063	1.812	0.085	0.054	1.570
CZE	-0.505	0.043	0.053	0.805	-0.028	0.036	-0.786
DEU	2.593	0.033	0.050	0.664	0.122	0.044	2.813
DNK	1.174	-0.017	0.040	-0.423	-0.051	0.036	-1.413
ESP	-3.900	-0.118	0.045	-2.619	-0.137	0.036	-3.796
FIN	-2.049	0.014	0.039	0.354	0.005	0.037	0.136
FRA	-1.465	0.035	0.046	0.760	-0.001	0.036	-0.028
GBR	-3.506	-0.245	0.039	-6.346	-0.214	0.032	-6.594
HUN	0.045	0.035	0.055	0.642	0.039	0.115	0.342
IRL	-1.647	0.036	0.048	0.752	0.035	0.041	0.874
ISL	-6.572	-0.108	0.027	-4.080	-0.155	0.027	-5.796
ITA	-3.933	-0.118	0.048	-2.426	-0.126	0.045	-2.831
JPN	-2.338	-0.106	0.073	-1.466	-0.279	0.073	-3.826
KOR	3.842	0.166	0.044	3.767	0.114	0.038	3.013
LUX	18.553	0.381	0.026	14.545	0.300	0.025	11.789
MEX	-5.154	-0.338	0.062	-5.452	-0.292	0.051	-5.690
NLD	-3.546	-0.193	0.051	-3.809	-0.157	0.040	-3.896
NOR	-2.155	-0.203	0.048	-4.222	-0.244	0.044	-5.547
NZL	-1.009	-0.086	0.044	-1.966	-0.117	0.041	-2.861
POL	2.982	0.172	0.062	2.751	0.144	0.055	2.638
PRT	0.520	0.122	0.071	1.723	0.082	0.059	1.390
RUS	-3.984	-0.174	0.067	-2.611	-0.203	0.059	-3.413
SWE	-0.261	-0.177	0.038	-4.589	-0.164	0.031	-5.308
USA	-1.189	-0.088	0.096	-0.924	-0.212	0.067	-3.173

## Different Models for Estimating Trends

SCIENCE	ORIGINAL TREND	MARGINAL TREND			CONDITIONAL TREND		
Country	Z-score	Unstandardized regression coefficient	S.E.	Z-score	Unstandardized regression coefficient	S.E.	Z-score
AUS	0.414	-0.061	0.044	-1.384	-0.078	0.035	-2.234
AUT	-5.457	-0.239	0.047	-5.131	-0.172	0.038	-4.570
BEL	3.564	0.185	0.050	3.690	0.152	0.044	3.421
CAN	-0.879	0.015	0.027	0.562	0.053	0.022	2.359
CHE	3.201	0.183	0.065	2.818	0.148	0.056	2.657
CZE	3.127	0.108	0.044	2.439	0.020	0.037	0.558
DEU	4.202	0.172	0.049	3.522	0.243	0.042	5.842
DNK	1.459	0.003	0.040	0.071	-0.031	0.033	-0.927
ESP	-2.924	-0.110	0.044	-2.483	-0.126	0.035	-3.607
FIN	4.014	0.040	0.031	1.295	0.028	0.029	0.967
FRA	3.856	0.149	0.047	3.154	0.119	0.040	2.963
GBR	-2.462	-0.123	0.038	-3.257	-0.113	0.031	-3.696
HUN	0.399	0.059	0.053	1.121	0.012	0.109	0.106
IRL	-0.556	-0.038	0.044	-0.863	-0.038	0.036	-1.074
ISL	-1.059	0.034	0.023	1.503	-0.017	0.024	-0.695
ITA	0.341	-0.019	0.046	-0.416	-0.026	0.043	-0.604
JPN	-2.025	-0.157	0.079	-2.001	-0.294	0.083	-3.551
KOR	-2.409	-0.135	0.050	-2.695	-0.223	0.048	-4.635
LUX	14.887	0.460	0.022	20.970	0.388	0.023	16.953
MEX	-5.735	-0.199	0.052	-3.833	-0.160	0.043	-3.754
NLD	-0.094	-0.084	0.052	-1.608	-0.062	0.044	-1.396
NOR	-4.579	-0.120	0.043	-2.797	-0.157	0.039	-3.989
NZL	-0.883	-0.067	0.039	-1.726	-0.096	0.037	-2.599
POL	1.823	0.161	0.061	2.645	0.144	0.054	2.676
PRT	2.505	0.140	0.059	2.380	0.096	0.051	1.897
RUS	3.588	0.199	0.073	2.735	0.168	0.067	2.512
SWE	-0.475	0.015	0.038	0.397	0.027	0.032	0.829
USA	-0.751	-0.041	0.089	-0.456	-0.158	0.063	-2.507

This document was created with Win2PDF available at <http://www.daneprairie.com>.  
The unregistered version of Win2PDF is for evaluation or non-commercial use only.