Maciej Jakubowski
Faculty of Economic Sciences, Warsaw University
Organisation for Economic Co-operation and Development
mjakubowski@uw.edu.pl

Artur Pokropek
Institute of Philosophy and Sociology of the Polish Academy of the Sciences
artur.pokropek@gmail.com

Reading achievement growth across countries[1]

INTRODUCTION

International surveys of students like PIRLS and PISA provide comparisons of learning achievement across participating countries. Although country rankings usually receive considerable attention, other results are far less discussed. While ranking position of a country is clearly important when assessing the overall potential of its students, this is less informative for education policy as these differences might be strongly affected by other characteristics than just the performance of school systems. In national assessments, school mean score is rarely taken as a measure of school efforts and the so-called value added scores are often considered as much more reliable indicators of school performance. We provide similar estimates to value-added scores but at the country level. We compare reading achievement in primary schools as measured by PIRLS to reading achievement of 15-year-olds in the PISA survey. We discuss various issues related to comparability of the results of these two surveys. We adjust achievement growth estimates for differences in the distribution of student background characteristics and for the difference in testing age across countries and surveys. We also provide a direct test of how changing a pool of test items would affect country rankings in respect to achievement growth captured by our methods. Finally, we compare achievement growth at different student performance levels using quantile regression approach. Generally, the results demonstrate remarkable consistency suggesting noticeable differences across countries in the reading achievement growth between primary and secondary education levels. The quantile regression evidence demonstrates also in which countries the low- or high-achieving students are progressing relatively slower or faster. These results can be used to assess the effectiveness of different school systems in their work with students between the age of 10 and 15.

[1] This research was supported by a financial grant from the Polish Ministry of Science and Higher Education and was administered by CASE - Center for Social and Economic Research.

PISA and PIRLS – commonalities and differences

This study analyzes publicly available datasets provided by organizers of PIRLS and PISA. Data as well as documentation are accessible online (for PIRLS http://timss.bc.edu; for PISA see www.pisa.oecd.org). The data used in this paper differ in two respects. Firstly, we re-standardize scores to the PISA scale and sample of countries participating both in PIRLS and PISA. This is explained in details below. Secondly, we use all data published in the datasets, even if they were not considered in the official publications (for example data from Netherlands in PISA 2000). Moreover, we separate data for England and Scotland as they participated in PIRLS 2001 independently and were published under "United Kingdom" in PISA 2000.

Both PIRLS 2001 and PISA 2000 aimed at measuring student achievement in reading in an internationally comparable way. While organization and people involved differ, the general notion is that these two studies have many things in common. They are based on similar methodologies, experts involved often work for both studies, and despite some visible differences their general goal is similarly stated. Both want to provide internationally comparable measures of what students can do in reading. The main difference is that PIRLS is conducted in primary schools while PISA measures achievement in secondary schools. That opens the possibility to compare how countries differ in achievement growth from primary to secondary education.

To start comparing results from these two surveys one needs to put them on a common scale. Performance scales in both surveys are derived from IRT models and standardized to have mean 500 and standard deviation 100 in a chosen group of countries. While in PIRLS this group comprises all participants, in PISA only OECD countries are considered when standardizing the scale. Both choices are arbitrary but we decided to re-standardize PIRLS scores to put them on the PISA scale, because the latter has a more meaningful interpretation with a mean of 500 among OECD countries as a commonly used benchmark. As not all OECD countries participated in PIRLS 2001 the scale has mean different than 500, namely, it is slightly lower in our sample of countries. However, the same benchmark of 500 as the OECD average can still be referred to.

Differences in sampling frames and testing framework seem to be crucial when thinking about direct comparison of PIRLS and PISA results. We address both in this paper. PIRLS surveys pupils in a grade with the highest number of 10-year-olds, which in most countries is a 4th grade. PISA surveys 15-year-olds regardless in which grade they are currently enrolled. This leads to higher variation in age for PIRLS and higher variation in grade for PISA. In our approach we adjust for differences in age distribution to address this issue. Adjustments for grade distribution are far more discussable as countries differ in grade retention or promotion policies and grade cannot be considered as given to a student. In fact, countries participating in PIRLS importantly differ in average age of students and within-country age distributions (see Table 1).

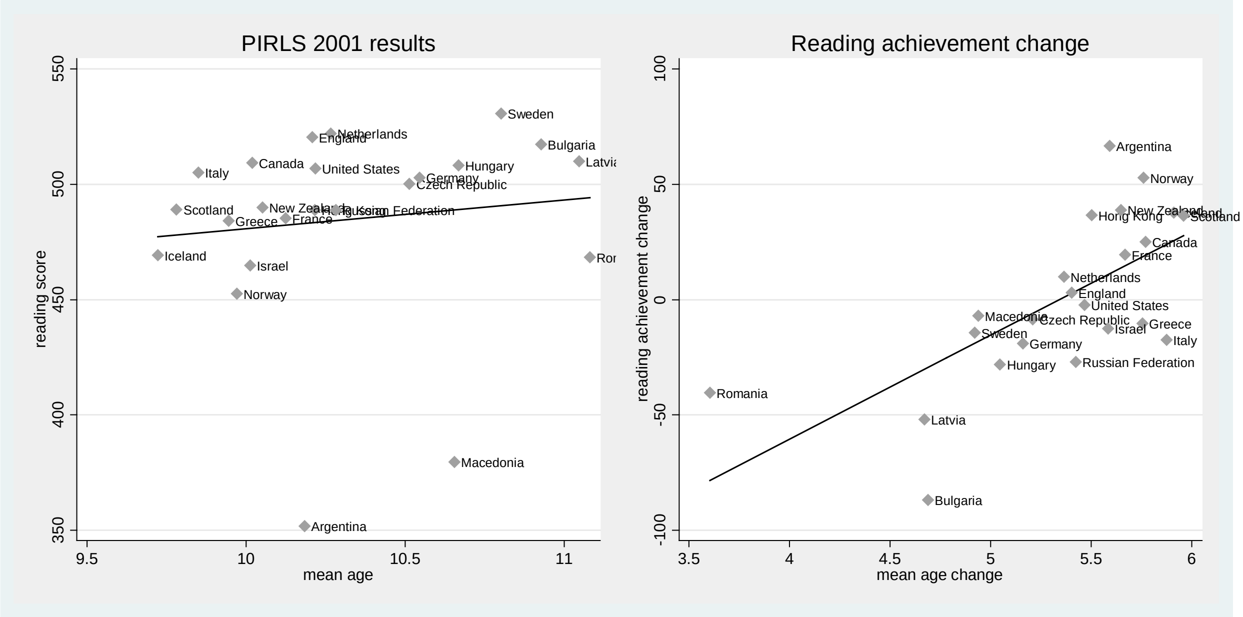
Table 1. Sample size, average reading performance and mean student age in countries participating in both PISA 2000 and PIRLS 2001.

Country	PIRLS 2001			PISA 2000		
	No of obs.	*Mean score*	*Mean age*	*No of obs.*	*Mean score*	*Mean age*
Argentina	3300	351.6	10.2	3983	418.3	15.8
Bulgaria	3460	517.4	10.9	4657	430.4	15.6
Canada	8253	509.4	10.0	29687	534.3	15.8
Czech Rep.	3022	500.2	10.5	5365	491.6	15.7
France	3538	485.3	10.1	4673	504.7	15.8
Germany	7633	503.0	10.5	5073	484.0	15.7
Greece	2494	484.1	9.9	4672	473.8	15.7
Hong Kong	5050	488.8	10.2	4405	525.5	15.7
Hungary	4666	508.2	10.7	4887	480.0	15.7
Iceland	3676	469.2	9.7	3372	506.9	15.6
Israel	3973	464.8	10.0	4498	452.2	15.6
Italy	3502	505.0	9.8	4984	487.5	15.7
Latvia	3019	509.9	11.0	3893	458.1	15.7
Netherlands	4112	522.1	10.3	2503	531.9	15.6
New Zealand	2488	490.0	10.1	3667	528.8	15.7
Norway	3459	452.4	10.0	4147	505.3	15.7
Romania	3625	468.3	11.1	4829	427.9	14.7
Russian Fed.	4093	488.8	10.3	6701	461.8	15.7
Sweden	6044	530.7	10.8	4416	516.3	15.7
Macedonia	3711	379.5	10.7	4510	372.5	15.6
United States	3763	506.8	10.2	3846	504.4	15.7
England	3156	520.4	10.2	4120	523.4	15.6
Scotland	2717	489.1	9.8	2371	525.6	15.7

Age differences affect mean score results and need to be taken into account when constructing achievement growth estimates. In PIRLS 2001, students in countries like Latvia or Romania were more than one year older than student in countries like Iceland, Italy or Scotland. Across all PISA countries the average age is very similar. In effect, the age difference between students tested in PIRLS 2001 and PISA 2000 importantly varies between countries. It is quite difficult to see how age differences affect comparability of mean scores in PIRLS, because cross-sectional nature of data doesn’t allow distinguishing between age effects and impacts of other difference between countries. However, the effect of mean age is much more visible when comparing outcomes over time within countries. In our case, country fixed characteristics are automatically taken into account because we look at achievement differences within countries. In this case, the impact of age differences is easier to detect.

Figure 1 demonstrates how testing age differences affect comparability of achievement measures. The left panel demonstrates that relation between mean age and reading scores is slightly positive when looking only at the PIRLS data. The right panel clearly shows that relation among the mean age difference between PIRLS and PISA and same difference in reading performance which is our basic measure of reading achievement growth is clearly positive. This simple evidence suggests that any achievement growth estimates taken from comparisons between PIRLS and PISA has to be adjusted for age effects. In this paper we propose a simple method which addresses this issue.

Figure 1. Relation between country average age and performance (left panel) and relation between change in country average age and performance from PIRLS to PISA (right panel).



Another critical difference is related to distinct assessment frameworks. While it is argued that PIRLS and reading literacy in PISA are closer to each other in test content than TIMSS and PISA assessments in mathematics and science, the frameworks were developed separately by distinct groups of experts working under different assumptions. Typically, PISA is considered as a test aim at measuring literacy needed in real-life situations and PIRLS or TIMSS are assessments more closely related to countries school curricula. Nevertheless, for reading testing domains in PIRLS and PISA seem to be similar in some parts. Careful analysis of test content and framework assumptions revealed many commonalities (Mullis et al., 2006, see Appendix C). We are aware of only one study which empirically addressed these issues using item level data (see Grisay, Gonzales and Monseur, 2009). Other studies discussed comparability of PIRLS and PISA results more generally (see Brown et al., 2005; Jakubowski, 2010).

We do not attempt to compare test contents in PIRLS and PISA, but aim at checking empirically how different pools of test items affect our outcomes. Our assumption is that if the two tests differ importantly in some content parts then taking smaller number of test items should visibly affect our results. More precisely, if the volatility of achievement growth estimates is relatively high when considering only parts of items pools from two surveys then content comparability would seem to be an issue. If countries' positions are stable despite the item pool taken for comparison then content differences can be considered as negligible. The details of the methodology are given in the following section.

METHODS

Dealing with complex survey design and missing data

PIRLS and PISA are complex surveys of student populations with multistage sampling designs that put additional burden on estimation. Non-response is also a problem in these surveys unless one wants to concentrate solely on performance data. Finally, both surveys use plausible values methodology which adds additional complications as all calculations has to be conducted separately for each of 5 plausible values and then combined into one final estimate. In this paper, we account for all these statistical requirements, however, not always following the solutions proposed in survey documentation.

All results presented here were obtained with survey weights provided in public PIRLS and PISA datasets. However, as we pool all the data into one dataset analyzed in one regression model we re-standardized weights to have the same sum for each country and survey. This way, each country and each survey contribute equally to final estimates.

The most demanding in this case is estimation of standard errors. PIRLS and PISA use different replication methods to adjust for complex survey design. Jackknife approach proposed in PIRLS and BRR method suggested by PISA are serving similar purposes but cannot be used together. Moreover, empirical checks and information provided in technical reports suggested that some information used to produce jackknife and BRR weights was not provided in public datasets for confidentiality reasons. Thus, it is not possible to reconstruct replication weights using a method preferred in one of the surveys. We decided to use another approach, the bootstrap method, which can be applied when only basic survey design information is provided. The bootstrap approach accounts for schools as primary sampling units and treats countries as survey strata, which probably give slightly overestimated standard errors. In other words, the confidence intervals provided in this paper are conservative in comparison to the original ones.

Finally, we had to address missing data issue. Many of our results are derived by adjustments based on background student characteristics. We decided to impute missing data on these characteristics not to invalidate mean performance estimates by dropping students who has missing value on background data. We used a multiple imputation model implemented in Stata procedure –ice- (Royston, 2004). The model was applied separately for each plausible value producing one imputed dataset with no missing data for each plausible value. The final estimates by aggregating the results of regressions conducted separately with each imputed dataset

Methods for achievement growth estimation

All estimates of achievement growth presented in this paper were obtained from linear regression with country dummy (0/1) variables. The regression was run on data from all countries and both surveys pooled in one dataset, giving equal weight to each country and survey. Country indicators were interacted with an indicator for PISA results and the coefficient on this interaction term denotes achievement growth, namely, a change of performance scores in this country from PIRLS to PISA. Several adjustments to this regression model were implemented addressing distinct needs.

Firstly, we provide results by subpopulations. The results for all students are accompanied with separate results for males and females, native students defined as those born in the country of the test and speaking at home the language of the test, and native males and females. Moreover, we adjust for the individual age effect in all above estimates. That is done by adding to regressions the best-fitting second degree fractional polynomial of student age, which captures the overall positive relation between age and student performance.

In the second method, we reweight the data to balance distribution of important background student characteristics across surveys. This follows a methodology proposed by Tarozzi (2007) and generally based on the idea close to the propensity score matching (Rosenbaum, Rubin, 1983). For different sets of background characteristics, a logit regression was applied to predict for each student a probability of being sampled for PISA 2000 having characteristics from a sample of PIRLS 2001. In other words, in the logit regression dependent variable was equal 1 for students sampled in PISA 2000 and 0 for students sampled in PIRLS 2001. The independent variables were students background characteristics which we would like to balance across PIRLS and PISA. From this model the probability of participating in PISA in respect to student background characteristics variables was predicted. Then PISA data were reweighted by the inverse of this probability to match the distribution of background characteristics in PIRLS. Finally, same as previously regression model was run on this reweighted data. Reweighted was conducted to adjust for % of immigrant students (defined as above), % of females, and % of students having different number of books at home and parents with different education levels.

Although regression analysis provides estimates of mean achievement growth across countries, one might be interested in looking at similar estimates across performance distribution. We provide such estimates by employing a quantile regression approach. We estimate the same regression models but on 10th, 25th, 75th and 90th percentiles of reading performance distribution. This way we provide estimates of achievement growth across low- and high-achieving students. Those can be used as evidence related to changes in inequality of student performance. For example, if the growth among high-achievers is bigger than among the low-achievers then one can conclude that variation in student scores increased in this country mainly by higher growth in the top of distribution. Such evidence can be useful when assessing distributional impact of education policies, when not only average achievement is considered but rather differences in achievement across low- and high-performing students are of interest.

Finally, we correct all the results for differences in mean age of tested students across PIRLS and PISA. Figure 1 demonstrates that such differences are strongly related to achievement change. We adjust for that by regressing the estimate of achievement growth on the difference in average age between PIRLS and PISA and its squared term. The residual from this regression is our mean age-adjusted measure of achievement growth. We apply this approach to all our estimates of achievement growth.

Random items draws

Serious bias in estimation of achievement growth may arise from differences in assessment frameworks in PIRLS and PISA. Even if the two surveys aim at measuring student achievement or literacy in the same domain, they define reading domain differently and consequently use different items pool. In this section, we describe our empirical attempts to assess how different choices of test items affect comparability of student outcomes from two surveys, more precisely, how they affect variability of achievement growth estimates.

In PIRLS 2001 and PISA 2000 reading was assessed with large pool of items (98 in PIRLS 2001 and 129 in PISA 2000) administered in balanced booklet design (for details technical reports for PIRLS 2001 and PISA 2000). This large number allows us to empirically test how changes in the pool of testing items affect final outcomes. If the two tests differ importantly in some content parts, then taking smaller number of test items should visibly affect the results. If countries’ positions are stable despite the item pool taken for comparison, then differences in test items pool can be considered as negligible for final estimates of achievement growth.

We employed a small simulation study to assess how achievement growth estimates can differ depending on choice of test items. In the first step, items were randomly sampled from both surveys. Each item was sampled independently of others with the same probability of selection equal to ½. Thus, on average half of the items were sampled from the item pool, but the actual number of test items differed across simulated samples (see Table A5 in Appendix). In the second step, we scaled student responses by the two-parameter IRT model (2PL) and assigned scores using expected-a-posteriori (EAP) method using PARSCALE 4.0 (see Muraki & Bock, 1997). No adjustments for booklets designs and no additional variables were used during the calibration process. Due to complex booklet designs some of specific sampled sets of items were unable to converge in calibration process because of too small number of shared items (these problems appear in PISA and not in PIRLS). Those sets were not taken into account in computations. Scores were standardized to have mean 500 and standard deviation 100 in 23 countries, which have participated both in PISA 2000 and in PIRLS 2001. In the last step, we computed average mean of PIRLS and PISA scores from the scaled student scores and subtracted them to obtain unconditional achievement growth estimates. These three steps were repeated 400 times to see how achievement growth estimates are affected by choice of test items.

RESULTS

Table 2 present results for unadjusted achievement growth estimates. First column present results for all students sampled in PIRLS 2001 and PISA 2000. Other columns present results for subpopulations: native students (those born in the country of the test), males and females, and native males and females. These results are not taking into account any differences between two surveys. They represent a simple difference between re-scaled scores from PIRLS and PISA. We will use them as a baseline for comparisons with adjusted achievement growth estimates. However, even from these data it is clear that achievement growth differs among subpopulations, quite importantly in some countries. For example, while students in Canada have relatively higher achievement growth this positive effect is much smaller for males than for females. The difference is even higher when looking at native students only. In other countries, such divergence is even more noticeable.

Table 2. Unadjusted achievement growth in population and subpopulations.

Country	All		Native		Males		Females		Native males		Native females	
	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Argentina	66.6	(9.0)	36.8	(9.2)	53.4	(10.0)	74.5	(9.8)	23.5	(10.5)	45.6	(9.9)
Bulgaria	-87.0	(7.8)	-87.3	(7.5)	-94.1	(8.7)	-77.6	(8.6)	-93.2	(8.5)	-79.3	(7.6)
Canada	25.0	(2.9)	16.1	(2.8)	19.7	(3.2)	30.0	(3.1)	8.9	(3.2)	23.6	(3.2)
Czech Rep.	-8.6	(6.0)	-12.4	(5.8)	-20.3	(6.6)	1.4	(6.6)	-24.8	(6.4)	-1.5	(6.3)
England	3.0	(6.0)	-7.8	(5.8)	5.0	(7.2)	1.8	(7.3)	-6.4	(7.0)	-8.1	(6.8)
France	19.4	(6.4)	18.7	(6.1)	11.3	(6.9)	26.3	(6.9)	10.5	(6.8)	25.7	(6.2)
Germany	-19.0	(5.9)	-20.8	(5.7)	-28.1	(6.3)	-9.7	(6.2)	-31.0	(6.7)	-10.1	(5.8)
Greece	-10.3	(6.8)	-9.7	(6.9)	-15.6	(8.0)	-5.0	(8.2)	-13.8	(8.0)	-5.9	(7.0)
Hong Kong	36.7	(6.2)	42.6	(6.3)	41.3	(7.6)	32.3	(7.5)	46.5	(7.5)	39.0	(6.3)

Hungary	-28.2	(6.3)	-36.2	(6.1)	-34.6	(6.7)	-21.2	(6.8)	-43.2	(6.8)	-28.4	(6.7)
Iceland	37.7	(4.5)	30.0	(4.5)	29.6	(5.2)	45.4	(5.2)	22.4	(5.4)	37.4	(4.9)
Israel	-12.6	(9.0)	-29.4	(8.5)	-7.9	(10.2)	-19.7	(10.2)	-25.3	(9.7)	-34.8	(9.2)
Italy	-17.6	(6.2)	-10.4	(6.0)	-31.7	(7.1)	-3.2	(7.0)	-23.5	(7.1)	1.1	(6.3)
Latvia	-51.9	(5.8)	-51.5	(6.4)	-65.8	(6.4)	-40.4	(6.2)	-70.0	(6.9)	-37.0	(6.8)
Macedonia	-7.0	(9.9)	5.4	(9.8)	-18.6	(10.3)	5.2	(10.5)	-5.0	(10.2)	15.3	(9.9)
Netherlands	9.8	(7.0)	15.8	(6.7)	4.8	(8.3)	15.3	(8.2)	9.7	(8.2)	22.3	(6.1)
New Zealand	38.8	(6.9)	36.7	(6.2)	32.4	(8.0)	44.8	(8.2)	31.6	(7.7)	40.6	(7.3)
Norway	52.8	(4.7)	52.4	(4.5)	44.7	(5.8)	60.9	(5.8)	45.0	(5.7)	59.5	(4.7)
Romania	-40.4	(6.9)	-43.5	(7.2)	-38.4	(7.7)	-42.9	(7.7)	-41.2	(8.1)	-46.3	(7.3)
Russian Fed.	-27.1	(5.9)	-31.4	(5.7)	-38.8	(6.2)	-15.7	(6.1)	-44.3	(5.9)	-18.8	(5.8)
Scotland	36.4	(6.1)	23.5	(6.8)	32.3	(6.5)	40.6	(6.6)	21.7	(7.7)	25.1	(7.8)
Sweden	-14.4	(4.3)	-16.4	(3.8)	-19.1	(4.9)	-9.7	(4.8)	-21.1	(4.6)	-11.8	(3.8)
USA	-2.4	(6.9)	-10.7	(6.4)	-5.6	(7.9)	0.1	(7.9)	-12.4	(7.8)	-9.9	(6.7)

We now turn to results obtained after controlling for individual student age effect on performance which are presented in the appendix (Table A1). Please note that while these estimates address somehow the issue of discrepancies in age distributions between PISA and PIRLS they do not account for major differences in average age across countries. Age was centered on survey mean, so it has mean 0 in PIRLS and PISA (weighting country data equally). Then it was modeled using a best-fitting polynomial of a second degree, which means that non-linear relation between age and performance was flexibly modeled. However, for majority of students the relation was positively sloped which reflects an intuition that older students of similar ability will perform better on the test just because of maturity effect. Nevertheless, the results are very similar to unadjusted estimates presented in Table 2.

Table A2 in the appendix presents results for achievement growth estimates obtained after reweighting. After reweighting, different sets of variables were balanced across PIRLS and PISA samples. Variables balanced in each reweighted sample are listed in columns headings. Thus, the first column balances student share of females and a share of students with an immigrant background (born outside the country of the test or speaking at home different language than the language of the test). The second column accounts for the distribution of socio-economic characteristics of students using two variables which provide comparable information on that in both PIRLS and PISA. The number of books at home and parental education were recoded to similar categories in both surveys and then balanced across two surveys. The estimates presented in the last column were adjusted for all these factors. They account for differences in the distribution of gender, immigrant status and family socio-economic background across PIRLS and PISA. Please note that estimates reweighed for socio-economic background are on average negative, which reflects the overall unbalance in these characteristics between PIRLS and PISA.

Finally, we re-estimated all above models accounting for differences in mean age across countries. That was done by regressing the achievement growth estimates on the difference in mean age across surveys for each country and its squared term. The residuals from this country-level regression were taken as the mean age-adjusted achievement growth estimates and are presented in tables below. In other words, these estimates are taking into account that country samples covered students of different age in PIRLS and PISA. In our view, they are much more reliable as cross-country comparisons of achievement growth as the mean-age difference between PIRLS and PISA is highly correlated with unadjusted achievement growth estimates.

Table 3. Mean age adjusted regression estimates for subpopulations

Country	Unconditional	Natives	Males	Females	Native males	Native females	Unconditional	Natives	Native males	Native females
	Not adjusted for student age					Adjusted for student age				
Argentina	55.6	30.5	49.0	57.4	23.7	33.4	52.5	27.5	20.8	30.2
Bulgaria	-49.5	-49.3	-50.0	-46.7	-47.5	-48.5	-52.8	-52.9	-51.0	-52.3
Canada	0.3	-2.5	1.0	0.1	-4.2	0.3	1.1	-1.8	-3.6	1.0
Czech Rep.	5.0	3.6	0.5	8.1	-1.0	7.1	6.8	5.3	0.8	8.8
England	4.9	-2.4	13.9	-2.9	6.2	-9.5	8.0	0.7	9.3	-6.5
France	2.8	7.4	1.0	3.9	5.2	8.9	3.4	7.8	5.6	9.5
Germany	-2.7	-2.3	-4.6	-0.4	-4.7	0.8	-1.6	-0.8	-3.3	2.5
Greece	-33.6	-27.1	-32.9	-33.7	-25.7	-28.2	-33.9	-27.5	-25.9	-28.8
Hong Kong	32.0	42.0	43.4	21.3	52.8	32.1	28.8	41.8	52.4	31.9
Hungary	-6.0	-12.3	-5.1	-5.9	-11.2	-12.0	-5.8	-12.0	-11.2	-11.5
Iceland	1.5	1.1	-1.3	4.6	-2.1	4.8	-0.2	-1.0	-4.1	2.8
Israel	-23.1	-35.2	-11.8	-36.2	-24.5	-46.5	-21.8	-34.0	-23.3	-45.5
Italy	-50.5	-36.5	-59.2	-40.9	-44.8	-28.9	-51.3	-37.5	-45.8	-29.9
Latvia	-13.8	-12.8	-21.1	-8.8	-23.7	-5.6	-18.0	-17.7	-28.9	-10.0
Macedonia	20.3	34.0	15.9	25.6	31.6	36.3	21.6	35.2	33.1	37.2
Netherlands	14.2	23.4	16.2	13.0	24.7	22.9	15.8	24.9	26.4	24.3
New Zealand	23.7	26.7	23.7	23.8	27.8	25.0	25.3	28.0	29.1	26.3
Norway	29.1	34.6	27.0	31.8	32.7	37.0	30.1	35.4	33.2	38.2
Romania	10.8	9.5	13.0	8.4	11.6	7.4	12.1	11.0	13.2	8.9
Russian Fed.	-26.4	-27.2	-31.1	-21.6	-32.8	-21.3	-26.7	-27.8	-33.4	-21.9
Scotland	-3.8	-9.1	-2.8	-4.0	-6.9	-10.8	-5.2	-10.6	-8.3	-12.3
Sweden	13.8	13.0	16.2	11.6	16.3	10.0	14.5	13.4	16.8	10.3
USA	-4.7	-9.1	-1.0	-8.6	-3.8	-14.9	-3.0	-7.4	-2.2	-13.0

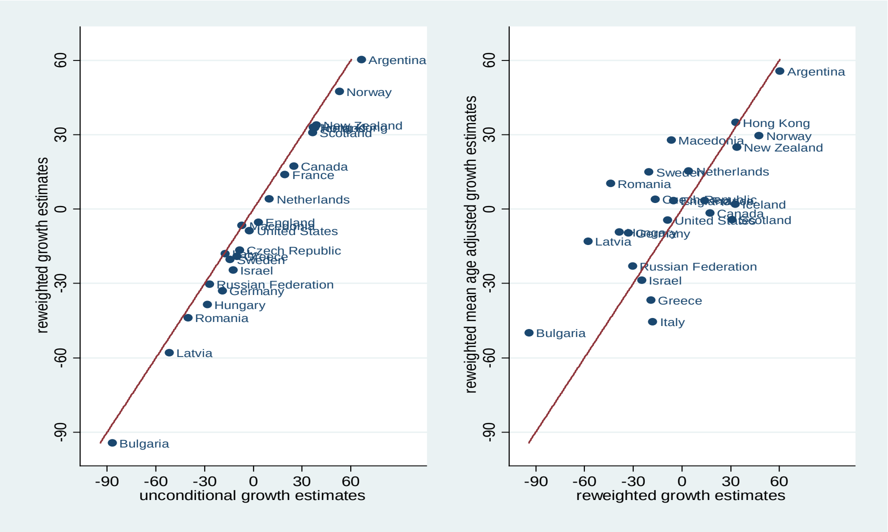
Table 4. Mean age adjusted reweighted regression estimates

Country	Data reweighed to balance the distribution in:		
	Gender and immigrant background	*number of books at home and parental education*	*Everything together*
Argentina	54.7	56.2	55.6
Bulgaria	-47.0	-52.9	-49.9
Canada	-0.2	-0.9	-1.6
Czech Rep.	4.2	5.0	4.0
England	4.7	3.8	3.3
France	2.4	3.6	3.3
Germany	-7.0	-5.6	-9.6
Greece	-36.7	-34.2	-36.7
Hong Kong	29.1	37.7	34.9
Hungary	-6.3	-8.6	-9.2
Iceland	1.4	2.5	1.9

Israel	-22.5	-29.3	-28.8
Italy	-47.0	-49.2	-45.5
Latvia	-14.9	-12.0	-13.0
Macedonia	25.5	22.9	27.9
Netherlands	13.6	16.0	15.3
New Zealand	25.9	22.9	24.9
Norway	28.7	29.6	29.6
Romania	10.4	10.8	10.3
Russian Fed.	-24.8	-24.2	-23.0
Scotland	-3.7	-4.3	-4.3
Sweden	12.3	16.7	15.0
United States	-2.9	-6.6	-4.5

Although all the estimates presented in table above differ somewhat in the magnitude of achievement growth they are at the same time relatively consistent. Definitely, the achievement growth estimates do not change noticeably after reweighting. Only controlling for the difference in average age across countries makes a highly noticeable difference for some countries. This is demonstrated on the graph below where on the left panel on can see correlation between unconditional estimates and those obtained after reweighting for all background characteristics, while the right panel presents the correlation between reweighted estimates before and after controlling for the mean age effect across countries. Clearly, the correlation on the right panel is much weaker demonstrating that adjusting for mean age makes considerable difference for final achievement growth estimates.

Figure 2. Correlation between unconditional and reweighted achievement growth estimates (left panel) and between reweighted estimates which were adjusted and non-adjusted for mean age effect (right panel).



The final estimates of achievement growth were obtained under similar assumptions but from quantile regression models. In other words, these are the same estimates but obtained for students of different ability levels. Because of space limitations, we present estimates for 10th, 25th, 75th and 90th percentiles only. Table 5 below presents results for the unconditional estimates while results for reweighted regressions and those adjusted mean-age differences are presented in the appendix. While standard errors are provided only for the simplest unconditional estimates, they should be similar across all models.

Table 5. Unconditional achievement growth by quantiles of performance.

	10th		*25th*		*75th*		*90th*	
Country	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Argentina	83.1	(11.7)	76.8	(11.8)	57.7	(9.7)	51.7	(10.4)
Bulgaria	-86.8	(10.8)	-95.1	(9.4)	-86.8	(8.3)	-81.1	(9.0)
Canada	18.6	(4.4)	20.9	(3.6)	28.2	(3.3)	28.3	(3.1)
Czech Rep.	-21.2	(8.7)	-15.4	(6.8)	-0.1	(6.3)	10.7	(6.7)
England	17.7	(9.6)	3.0	(7.7)	-1.2	(6.7)	-3.1	(7.5)
France	12.5	(8.4)	13.9	(7.8)	25.0	(6.2)	21.3	(6.2)
Germany	-52.4	(8.7)	-32.4	(7.5)	-0.3	(5.8)	13.3	(5.3)
Greece	-20.6	(10.4)	-15.3	(8.9)	-6.6	(6.6)	-2.5	(6.9)
Hong Kong	29.5	(10.5)	35.0	(7.7)	40.7	(5.3)	40.0	(5.3)
Hungary	-44.9	(8.4)	-41.5	(7.6)	-17.1	(7.2)	-11.6	(6.8)
Iceland	41.0	(7.2)	37.5	(5.9)	37.7	(4.9)	35.3	(5.6)
Israel	4.1	(15.4)	-11.3	(12.1)	-16.5	(7.8)	-18.2	(7.7)
Italy	-18.0	(9.7)	-20.0	(7.8)	-15.5	(6.4)	-12.7	(6.4)
Latvia	-84.0	(8.1)	-70.1	(7.5)	-32.7	(6.4)	-19.8	(7.2)
Macedonia	51.2	(12.0)	20.8	(13.5)	-36.8	(9.7)	-48.7	(8.7)
Netherlands	-20.3	(12.6)	-0.3	(10.2)	24.4	(5.9)	26.5	(6.4)
New Zealand	55.4	(12.0)	41.6	(9.3)	34.9	(7.0)	30.0	(7.1)
Norway	50.5	(8.6)	49.5	(6.8)	55.1	(4.8)	55.8	(4.5)
Romania	-21.6	(11.0)	-41.2	(10.9)	-48.5	(7.4)	-46.6	(7.3)
Russian Fed.	-40.8	(9.0)	-38.4	(6.9)	-20.5	(6.0)	-11.2	(6.3)
Scotland	46.9	(8.8)	36.6	(7.2)	33.9	(6.0)	28.3	(7.1)
Sweden	-32.2	(6.9)	-24.1	(5.4)	-4.6	(4.4)	-2.2	(4.6)
United States	-2.6	(10.5)	-7.7	(8.9)	-3.8	(6.4)	5.0	(6.7)

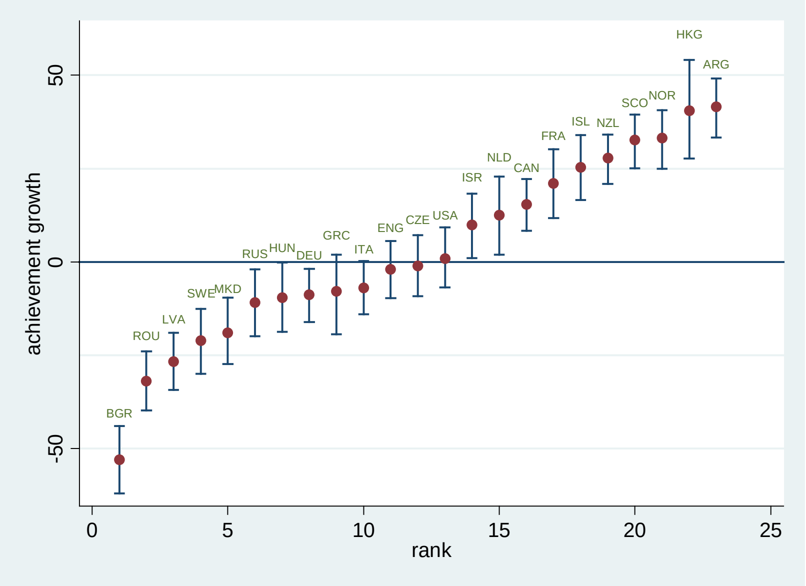
The results by level of performance reveal very interesting patterns in achievement growth across countries. Generally we can classify countries into three groups. Countries in the first group, for example Iceland or the United States, have similar achievement growth among students at all performance levels. In the second group achievement growth is more positive for low achieving students and smaller for top-achievers. This pattern is clearly visible in countries like New Zealand or Macedonia. From this evidence one can conclude that in these countries performance gaps are decreasing with student age in comparison to other countries. The third group includes countries like Netherlands or Latvia where the growth is higher for top-achievers than for low-achievers. In this group the gap between high- and low-achievers increases over time. One could note that among countries in the last group many have early tracking system where students are separated into different school programmes before the age of 15.

At the end, we present results for the simulation study which deals with the impact of item choice on achievement growth estimates. 400 simulated results were combined and simple statistics were computed. Detailed results are shown in Appendix A. Means, standard deviations (which may be treated as standard errors), and 5th and 95th percentiles were computed for each country. Table A6 in Appendix present these results for mean scores for each country as well as for the difference between them which is the simplest measure of achievement growth.

Results indicate that estimates based on different sets of items are relatively stable. Median standard deviation in mean PIRLS achievement estimators among all countries is 3.9 and median standard deviation for PISA mean achievement estimators is 3.2, which is close to the sampling error. For some countries, however, variation of estimates is higher. For example, Hong Kong has standard deviation of 7.5 in PIRLS and of 4.0 in PISA. This suggests that results for Hong Kong might be relatively more affected by the choice of items. On the other hand, for both PIRLS and PISA in Canada standard deviations were relatively low (2.4 in PISA and 3.2 in PIRLS), similarly for Scotland (3.0 in PISA and 2.9 in PIRLS).

Estimates of achievement growth demonstrate similar stability when using only randomly taken subsamples of items. Figure 3 demonstrates simulation results graphically (see Table A6 for detailed results). Points in graph represent the mean value of achievement growth estimates from 400 samples and intervals are based on empirical simulation distributions (the upper bound is 95th percentile of simulated estimates and the lower bound is the 5th percentile). Confidence intervals for 10 countries are clearly above 0 which indicates that despite the choice of items the estimate of achievement growth would be positive for them. For 8 countries we can say that achievement growth is relatively lower than in other countries. 5 countries are close to each other and their confidence intervals contain 0, so no clear conclusion on their achievement growth is possible. On the graph we can also clearly see that achievement growth estimators for Hong Kong are more volatile as one may expect from volatility of mean score estimates. Generally, this figure confirms that achievement growth estimates are quite stable. The ranking of countries would not change importantly with different sets of items considered for assessment.

Figure 3. Achievement growth: results from random items draws obtained from 400 simulations with random choice of test items.



SUMMARY

This paper provides internationally comparable results on reading achievement growth between the age of 10 and 15. The estimates were obtained by comparing individual student results from PIRLS 2001 and PISA 2000. These are two international surveys which while differ in some assumptions still provide comparable assessment tests in our view. We test this presumption empirically by employing a simulation study where achievement growth estimates were obtained from randomly taken test items. This exercise suggests that taking different sets of test items would provide us with similar conclusions that those based on full PIRLS and PISA results. In our view, that demonstrates robustness of our findings to small changes in assessment tools.

In the paper we provide results for subpopulations by gender and the immigrant background. We also provide results adjusted for differences in the distribution of important student characteristics like gender, immigrant background, parents education and the number of books at home. The results show remarkable consistency suggesting that differences in target populations and students sampling do not invalidate direct comparisons.

We found that the only thing which might invalidate direct comparisons is the difference in mean testing age across countries and between surveys. In some countries, students tested in PIRLS are much younger or older than the survey average, while in PISA students are of very similar age in all countries. That not only affects comparisons of PIRLS results but even more the estimates of achievement growth. Simply, in some countries students tested between PISA and PIRLS have one year less for educating themselves than in others. That has to affect the magnitude of achievement growth and we proposed a simple method to account for that. This kind adjustment makes a noticeable impact on our estimates. In the paper we provide full sets of results with mean age adjustments which can be taken as more reliable for between-country comparisons.

Finally, we provide evidence on the reading achievement growth across performance levels. These results are very interesting in our view as they show heterogonous is student progress across and within countries. These results are also much reliable when taken as relative measures of achievement growth within countries, because the differences in achievement growth between groups of students within the same country are less affected by discrepancies in testing framework and differences in other characteristics between countries. From these results, it is clear that in some countries low-performing students achieved less than high-performers in the period between age of 10 and 15, while in other countries it is the opposite. There are only few countries where achievement growth is similar across all performance levels. These results are of interest for countries looking for an internationally comparable way to assess the distributional effects of their policies. In fact, the results can be directly interpreted as demonstrating in which countries performance inequality is increasing or decreasing in the analyzed period.

REFERENCES

Brown G., Micklewright J., Schnepf S., Waldmann R., (2005). “Cross-National Surveys of Learning Achievement: How Robust are the Findings?”. Southampton Statistical Sciences Research Institute, S3RI Applications and Policy Working Papers, A05/05

Grisay A., Gonzalez E., Monseur Ch., (2009) "Equivalence of item difficulties across national versions of the PIRLs and PISA reading assessments", in: Davier M., Hastedt D. (eds.) “IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, Volume 2.”

Jakubowski M., (2010), “Institutional tracking and achievement growth. Exploring difference-in-differences approach to PIRLS, TIMSS and PISA data”. In: Dronkers J. (Ed.) “Quality and Inequality of Education. Cross-National Perspectives”. Springer.

Martin, M., Mullis I., Kennedy A. (Eds.) (2003), “PIRLS 2001 Technical Report”, Chestnut Hill, MA: Boston College.

Mullis I., Kennedy M., Martin M., Sainsbury M., (2006) “PIRLS 2006 Assessment Framework and Specifications”, International Study Center, Lynch School of Education, Boston College

Muraki E., Bock R. D. (1997). “*PARSCALE: IRT item analysis and test scoring for rating-scale data”.* Scientific Software International.

OECD, (2002). “*PISA 2000 Technical Report”,* OECD, Paris

Rosenbaum, P.R. and Rubin, D.B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", Biometrika 70, 1, 41-55.

Royston P. (2004). Multiple imputation of missing values. Stata Journal 4(3):227-241.

Tarozzi, A. (2007). “Calculating Comparable Statistics from Incomparable Surveys, with an Application to Poverty in India.” Journal of Business and Economic Statistics 25(3): 314-336

Appendix. Additional results

Table A1. Achievement growth adjusted for the effect of student age.

Country	All		Natives		Native males		Native females	
	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Argentina	60.4	(8.9)	30.4	(9.1)	17.1	(10.4)	38.8	(9.8)
Bulgaria	-93.1	(7.7)	-92.7	(7.4)	-97.8	(8.5)	-84.6	(7.6)
Canada	22.8	(2.9)	13.3	(2.8)	5.8	(3.2)	20.6	(3.3)
Czech Republic	-10.2	(5.9)	-13.8	(5.7)	-25.9	(6.5)	-2.9	(6.2)
England	2.8	(6.0)	-8.1	(5.8)	-6.7	(7.0)	-8.4	(6.7)
France	17.0	(6.5)	15.7	(6.1)	7.2	(6.8)	22.6	(6.1)
Germany	-21.2	(5.8)	-22.2	(5.7)	-32.4	(6.7)	-11.5	(5.8)
Greece	-13.4	(6.9)	-13.6	(7.0)	-17.7	(8.1)	-10.2	(7.0)
Hong Kong	30.2	(6.1)	39.0	(6.2)	42.6	(7.6)	35.3	(6.4)
Hungary	-31.4	(6.2)	-38.7	(6.1)	-45.6	(6.7)	-30.5	(6.7)
Iceland	33.4	(4.6)	24.6	(4.6)	16.6	(5.5)	31.8	(5.0)
Israel	-14.5	(9.1)	-31.7	(8.4)	-27.6	(9.7)	-37.4	(9.1)
Italy	-20.9	(6.2)	-14.9	(6.0)	-28.3	(7.1)	-3.6	(6.3)
Latvia	-58.8	(5.7)	-58.0	(6.4)	-76.1	(7.0)	-42.9	(6.8)
Macedonia	-9.0	(9.9)	4.1	(9.7)	-5.7	(10.2)	13.8	(9.7)
Netherlands	8.1	(6.9)	14.0	(6.7)	8.1	(8.1)	20.3	(6.0)
New Zealand	37.3	(7.0)	34.5	(6.0)	29.3	(7.8)	38.3	(7.2)
Norway	50.9	(4.8)	49.7	(4.7)	41.7	(5.7)	57.1	(4.8)
Romania	-37.2	(7.2)	-37.6	(7.3)	-33.8	(8.2)	-39.8	(7.6)
Russian Federation	-30.6	(5.9)	-35.4	(5.6)	-48.3	(5.9)	-22.9	(5.9)
Scotland	32.6	(6.1)	18.7	(6.8)	16.5	(7.7)	20.0	(7.8)
Sweden	-16.9	(4.3)	-18.4	(3.8)	-22.6	(4.5)	-13.8	(3.9)
United States	-4.0	(6.9)	-12.4	(6.3)	-14.2	(7.8)	-11.6	(6.7)

Table A2. Reweighted achievement growth estimates.

Country	Data reweighed to balance the distribution in:		
	Gender and immigrant background	*Number of books at home and parental education*	*All characteristics together*
Argentina	66.2	60.5	60.4
Bulgaria	-84.2	-97.6	-94.3
Canada	24.9	17.5	17.3
Czech Rep.	-9.0	-16.0	-16.5
England	3.3	-5.2	-5.3
France	19.6	13.8	13.9
Germany	-22.9	-29.3	-32.9

Greece	-12.8	-17.1	-19.1
Hong Kong	34.2	35.4	33.1
Hungary	-28.2	-38.4	-38.5
Iceland	38.1	33.1	32.9
Israel	-11.6	-25.6	-24.6
Italy	-13.5	-22.1	-18.0
Latvia	-52.7	-57.4	-57.9
Macedonia	-1.5	-12.0	-6.5
Netherlands	9.7	4.4	4.2
New Zealand	41.5	31.4	33.9
Norway	52.9	47.1	47.5
Romania	-40.8	-43.7	-43.9
Russian Fed.	-25.0	-31.9	-30.2
Scotland	37.1	30.5	30.9
Sweden	-15.6	-19.0	-20.3
United States	-0.2	-11.3	-8.8

Table A3. Reweighted achievement growth estimates by quantiles of performance.

Country	*10th*	*25th*	*75th*	*90th*
Argentina	82.0	75.9	57.6	51.6
Bulgaria	-81.3	-91.2	-84.8	-80.3
Canada	17.6	20.4	28.6	27.9
Czech Republic	-22.0	-16.0	-0.3	10.8
England	16.4	2.9	0.0	-1.4
France	11.7	13.8	25.8	22.1
Germany	-59.9	-38.4	-2.9	11.2
Greece	-24.2	-17.4	-8.3	-4.2
Hong Kong	27.2	32.5	38.4	38.3
Hungary	-45.0	-42.0	-16.7	-11.4
Iceland	40.1	37.3	39.0	36.9
Israel	3.8	-9.9	-15.1	-18.0
Italy	-13.7	-14.8	-12.4	-10.4
Latvia	-83.9	-69.4	-34.6	-23.2
Macedonia	56.5	27.5	-32.1	-45.1
Netherlands	-22.7	-0.5	25.2	27.6
New Zealand	59.1	44.8	36.9	32.9
Norway	51.3	49.8	55.2	55.8
Romania	-22.4	-41.6	-48.8	-46.5
Russian Federation	-38.8	-36.2	-18.3	-9.4
Scotland	47.4	36.8	35.2	29.7
Sweden	-34.4	-25.3	-6.0	-2.9
United States	2.2	-5.4	-2.2	7.1

Table A4. Unconditional and reweighted achievement growth estimates by quantiles of performance adjusted for mean age difference between PIRLS and PISA.

Country	Unconditional				Reweighted			
	10th	*25th*	*75th*	*90th*	*10th*	*25th*	*75th*	*90th*
Argentina	73.9	68.7	44.5	37.5	72.8	67.3	43.7	36.8
Bulgaria	-43.3	-51.6	-53.5	-53.6	-38.0	-48.2	-51.7	-52.6
Canada	-7.6	-2.3	3.1	3.9	-8.6	-3.1	2.6	2.7
Czech Republic	-1.3	3.2	9.2	16.1	-2.2	2.1	8.6	16.0
England	24.1	9.0	-2.8	-7.3	22.7	8.4	-2.1	-6.0
France	-3.7	-0.5	6.8	2.9	-4.5	-1.0	6.9	2.9
Germany	-29.5	-11.0	11.5	20.9	-37.1	-17.4	8.6	18.6
Greece	-45.3	-37.1	-30.6	-25.9	-48.9	-39.5	-33.2	-28.4
Hong Kong	28.1	33.8	33.2	30.6	25.7	30.9	30.3	28.4
Hungary	-15.7	-13.6	0.4	1.2	-15.9	-14.7	0.6	1.4
Iceland	0.0	1.6	2.6	2.6	-1.0	1.0	2.8	3.1
Israel	-4.4	-18.8	-29.2	-32.0	-4.8	-17.8	-28.5	-32.4
Italy	-54.9	-52.3	-47.9	-43.1	-50.6	-47.5	-45.7	-41.8
Latvia	-40.0	-25.9	1.2	8.4	-40.1	-25.7	-0.7	5.2
Macedonia	85.7	54.0	-14.1	-31.2	90.8	60.1	-9.7	-27.6
Netherlands	-11.0	8.4	25.1	24.3	-13.5	7.6	25.4	25.1
New Zealand	41.1	29.0	18.1	12.7	44.8	31.8	19.3	14.9
Norway	25.5	27.4	30.7	32.1	26.2	27.3	30.0	31.2
Romania	11.9	11.9	10.1	9.9	11.0	11.0	9.9	10.1
Russian Federation	-35.8	-33.8	-23.2	-16.4	-34.0	-32.1	-21.6	-15.0
Scotland	0.5	-3.8	-4.8	-7.4	1.0	-4.1	-4.5	-7.1
Sweden	3.0	9.9	18.8	16.1	0.7	8.2	17.3	15.4
United States	-1.2	-6.3	-9.2	-2.5	3.6	-4.5	-8.2	-0.8

Table A5. Number of items in different percentiles of sampled sets of items

Percentiles of sampled sets of items								
1%	5%	10%	25%	50%	75%	90%	95%	99%
PISA (number of items)								
52	55	57	61	65	68	72	73	77
PIRLS (number of items)								
38	41	43	46	50	53	56	57	60

Table A6. Results for simulations of different item pools. 400 Replications

Country	PIRLS 2001				PISA 2000				Achievement growth (PISA-PIRLS)			
	mean	sd	p5	p95	mean	sd	p5	p95	mean	sd	p5	p95
Argentina	426.4	3.9	420.1	472.5	467.9	2.9	462.9	472.5	41.5	4.8	33.3	49.2
Bulgaria	526.5	3.6	520.6	480.1	473.5	3.7	467.4	480.1	-53.0	5.3	-62.0	-43.9
Canada	506.3	3.2	501.4	526.0	521.8	2.4	517.6	526.0	15.5	4.1	8.3	22.2
Czech Rep.	508.3	3.5	502.2	513.4	507.2	3.7	501.6	513.4	-1.1	5.3	-9.2	7.1
England	522.4	3.5	516.5	525.0	520.4	3.1	515.0	525.0	-1.9	4.7	-9.8	5.6
France	501.3	4.5	494.0	527.2	522.2	3.0	516.6	527.2	21.0	5.4	11.7	30.1
Germany	516.5	3.5	510.8	512.6	507.7	2.9	503.1	512.6	-8.8	4.5	-16.1	-1.8
Greece	500.5	4.9	492.5	498.2	492.7	3.6	486.7	498.2	-7.8	6.3	-19.4	2.0
Hong Kong	487.8	7.5	475.5	534.6	528.3	4.0	521.3	534.6	40.5	8.6	27.7	54.1
Hungary	510.9	4.8	503.1	507.0	501.3	3.4	496.0	507.0	-9.6	5.7	-18.7	-0.2
Iceland	486.4	4.4	478.9	516.3	511.8	2.8	506.8	516.3	25.4	5.2	16.6	33.9
Israel	483.7	4.1	477.5	499.9	493.6	3.6	488.0	499.9	10.0	5.4	1.1	18.2
Italy	515.6	3.4	510.4	513.6	508.7	2.9	504.0	513.6	-6.9	4.4	-14.0	0.3
Latvia	506.5	3.9	499.2	484.5	479.8	2.9	475.3	484.5	-26.6	4.8	-34.3	-19.0
Macedonia	429.7	3.8	422.9	417.0	410.8	3.5	405.9	417.0	-18.9	5.0	-27.4	-9.6
Netherlands	510.6	4.8	503.2	530.6	523.1	4.2	516.2	530.6	12.5	6.5	1.9	22.8
New Zealand	498.7	3.1	493.7	530.8	526.5	2.9	521.7	530.8	27.8	4.0	20.9	34.0
Norway	482.2	4.2	475.2	520.2	515.3	2.9	510.3	520.2	33.2	4.9	25.0	40.7
Romania	483.4	4.0	476.8	457.1	451.4	3.6	445.6	457.1	-32.0	5.2	-39.8	-23.9
Russian Fed.	496.7	4.6	488.9	492.7	485.8	3.8	478.8	492.7	-10.9	5.5	-20.0	-1.9
Scotland	494.9	3.0	490.2	532.1	527.5	2.9	523.0	532.1	32.6	4.3	25.1	39.4
Sweden	543.0	4.5	535.5	527.7	522.0	3.2	516.8	527.7	-21.1	5.4	-30.0	-12.6
United States	499.4	3.6	493.2	506.0	500.3	3.2	495.0	506.0	1.0	4.8	-6.8	9.3