On the degree of comparability in trend studies as a function of differences in age and schooling

Monica Rosén, University of Gothenburg, Sweden, monica.rosen@ped.gu.se
Rolf Strietholt, Dortmund University of Technology, Germany, strietholt@ifs.tu-dortmund.de

Abstract

Linking IEA-studies on reading literacy of 9-10 year-old students via IRT-technique provides an extensive source for trend analyses from 1970 to 2006. Comparison derives from the assumption of having comparable groups in respective studies and countries. Regarding this, students' age and time of schooling play a crucial role since the sub-samples cover students that differ in terms of age, grade and test months. Therefore, the comparability can be considered as a function of differences in age and schooling. The main purpose of our analysis is highlight issues of study design that appear in trend analysis and affect the comparability. We discuss miscellaneous strategies to handle such differences and address limitations of these approaches.

Keywords: *age effect, comparative study, reading literacy, schooling effect*

Background and Study Purpose

Starting in the 1970s, the IEA launched studies on reading achievement of 9-10 year-old students (Thorndike, 1973; Elley, 1994; Martin, Mullis, Gonzalez, & Kennedy, 2003; Mullis, Martin, Gonzales, & Foy, 2003; Mullis, Martin, Kennedy, & Foy, 2007). Links between the studies enable measures of trends in reading literacy (Rosén, 2006; Rosén & Strietholt, 2010c). This is of great interest in educational research, since effects of educational reforms and other societal changes do typically not show until they are fairly well implemented. Unlike cross-sectional studies, trend studies make development visible over time. However, a number of methodological challenges adhere to such analyses. Therefore, it is particularly important to know in how far different methodological approaches affect the results and which approaches are most suitable. For instance, since the tests are not the identical in studies the reading achievement measures have to be linked properly, e.g. via item response theory (IRT). Also, possibilities to draw causal inference from non-experimental data are subject to conditions like the plausibility of alternative explanations for observed change (Shadish, Cook, & Campbell, 2002). In the same vein, the comparability of the data derives from the assumption of having comparable groups of students in the respective studies and countries. In this respect, students' age and time of schooling play a crucial role. Apart from the change in the IEA sampling-design from an age-based sample to an age-based sample of grades as a basic principle, apparently minor differences, such as the month of testing are often neglected. This is in so far important as a

significant amount of research has shown independent effects of both maturing and schooling on students' abilities (Cahan & Cohen, 1989; Ceci, 1991; e.g. Cliffordson, 2008; e.g. Cliffordson & Gustafsson, 2008; Luyten, 2006).

In the following section we first overview the sampling framework of the studies in our analysis and point out issues that affect students' age as well as time of schooling. Thereafter, we give a review on the significance of age and schooling on students' ability and achievement. The section closes with alternative strategies to deal with differences in age and schooling.

Sampling frameworks and data collection in IEA reading studies

Our analysis includes data from five IEA studies on reading literacy. The definition of the target population, which was defined by age and/or grade, has changed over time and, therefore, the samples in the different studies are not entirely comparable from the outset. Additionally, the month in which testing was conducted in the respective studies and countries have not been the same in all studies. Therefore, students spend more time in school in some samples than in others. This following section is divided into three parts that address these differences in grade, age and test month.

Grade Differences

In the 1970s Six Subject Survey (SSS) students at the age of 10 years were sampled. As a consequence, the data comprise students from more than one grade in the respective countries. It is crucial to note that the sampling within grades was not uniformly distributed; while students were relatively young in the upper grades, they were elder in the lower grades. Mainly, the data contains grade 3, 4, and 5 students, but this varies between countries. The sampling framework in 1991s Reading Literacy Study (RL) and the following studies was changed to grade-based sampling strategies. That means that the target population was first defined by age, e.g. the countries were to select the upper of the two adjacent grades which held most 9-year olds at the time of testing, but finally each country selected one grade. Mostly, this was grade 3 or 4. In the 2001s repetition of the Reading Literacy Study, comparability to the data of 1991 was a declared goal and, therefore, each country defined the target population the same way as before. In the more recent Progress in International Reading Literacy Studies (PIRLS) the definition of the target grade was the grade that represents four years of schooling, counting from the first year of ISCED Level 1.

Table 1 provides an overview of the samples in our study. With respect to grade, comparability between countries has not been achieved before all countries selected grade 4 students in PIRLS. However, SSS data contains grade 4 students from all countries although the sample size is rather small for Italy (n=159). For the above-mentioned reasons, some limitations come with the 1970s samples. In RL and PIRLS Italy and USA have grade 4 samples. Hungary and Sweden selected grade

3 in RL but grade 4 in PIRLS. Therefore, grade 4 data is available for Italy and the USA for all years but interrupted in 1991 for Hungary and Sweden. Note that Sweden expanded the international design in PIRLS 2001 with an additional grade 3 sample.

Table 1: Countries' sample sizes by study and grade

Study	Country	Number of students by grade				
		3	4	5	5 <	Total
SSS 1970	Hungary	–	3423	1396	5	4824
	Italy	5	159	4220	1	4385
	Sweden	899	1032	–	–	1931
	USA	83	1675	3494	17	5269
RL 1991	Hungary	3009	–	–	–	3009
	Italy	–	2221	–	–	2221
	Sweden	4297	–	–	–	4297
	USA		6433	–	–	6433
RL 2001	Hungary	4707	–	–	–	4707
	Italy	–	1590	–	–	1590
	Sweden	5361	–	–	–	5361
	USA	–	1826	–	–	1826
PIRLS 2001	Hungary	–	4666	–	–	4666
	Italy	–	3502	–	–	3502
	Sweden	5271	6044	–	–	11315
	USA	–	3763	–	–	3763
PIRLS 2006	Hungary	–	4068	–	–	4068
	Italy	–	3581	–	–	3581
	Sweden	–	4394	–	–	4394
	USA	–	5190	–	–	5190
Total		**23632**	**53567**	**9110**	**23**	**86332**

Differences in Test Month

Another source affecting comparability between countries and over time is related to the amount of schooling. Similar to students' grade, apparently minor differences in the month of testing also affect comparability but are often neglected. It seems plausible that students who are tested later tend to have higher scores in the reading test since they have extra time of schooling and opportunity to learn. And, therefore, even if comparisons restrain on a particular grade, changes in students' reading scores over time might rather reflect differences in the month of testing than an actual change. However, records of the test month did not exist in the datasets before 2001, so no information on test month is available for the 1970s SSS and 1991s RL study in the datasets.

In SSS the data collection was carried out at the end of the spring semester (Hansson, 1975).

According to Elley, the 1991s RL "tests and questionnaires were administrated (…) in the eighth month of the school year 1990-1991" (1992, p. 2). But, as mentioned above, before 2001the datasets provide no information weather the data collection actually adhered rigidly to this guideline. The trend report on the 2001 repetition outlines that the researchers were well aware that the month in which testing takes place might affect the results. Here it says that "[t]o ensure comparability over time, the 2001 data collection was scheduled in each country for the same time of year, as in 1991" (Martin, et al., 2003, p. 10). In PIRLS, testing was conducted toward the end of the school year, according to the reports most often in April to June of 2001 and 2006 (Mullis, et al., 2003; Mullis, et al., 2007).

Table 2 lists the test months for the respective countries and studies. From a within-country-trend-perspective, all samples from RL 2001 and PIRLS 2001 are comparable with respect to test month as both studies were set up and conducted simultaneously. As mentioned above, for the sake of comparability the time frame for the 2001s RL study was supposed to be the same as it was in 1991 in the respective countries. Assuming that this is true, comparability is also achieved for these two studies. However, the 1991s data lack information on test month and, therefore, this cannot be verified by the datasets. In PIRLS 2006 some differences in the test month can be observed. In Hungary and Italy testing took place one respectively two month earlier compared to the data collection in 2001. In Sweden, data was collected in April 2001 but spread on April and May in 2006. Noteworthy, in the USA the time frame span about half a year in 2006 and no information on this anomaly is available in the international report.

Table 2: Test month by study and country

Study	Hungary		Italy		Sweden		USA	
	month	**n**	**month**	**n**	**month**	**n**	**month**	**n**
SSS 1970	no information							
RL 1991	8th month of the school year							
RL 2001	-	-	-	-	April	5361	April	693
	May	4707	May	1557	-	-	May	1130
	-	-	June	33	-	-	-	-
PIRLS 2001					April	11299	April	1392
	May	4666	May	3502	May	16	May	2371
PIRLS 2006	-	-	-	-	-	-	January	1504
	-	-	-	-	-	-	February	1620
	March	33	March	3581	March	61	March	758
	April	4035	-	-	April	2765	April	483
	-	-	-	-	May	1568	May	684
	-	-	-	-	-	-	June	140

Age Differences

The importance of students' grade is evident since schools are probably the main educational institution where formal education takes place. However, apart from this, students also develop skills and abilities in other settings and due to their maturation. We used the term "schooling" to denote the amount of learning in school whereas "age" comprises out-of-school learning activities and student's maturation.

The mean age in month in the respective studies and countries is listed in table 3. In the most recent PIRLS 2006 study, all countries selected grade 4 as the target population. Italian students were 116 months old on average, whereas Swedish students' average age was 130 months. Such large age differences are mainly due to different policies and practice regarding the age of entry to primary school in the respective countries. However, some differences might also be due to different test months. In this regard, for instance, the age difference in the Italian PIRLS samples from 2001 (mean age = 118 months) and 2006 (mean age = 116 months) is 2 months, which isin line with the test months: the 2001 data were collected in May, the 2006 data in March (see table 2). A third source for age differences is the grade. In PIRLS 2001, Sweden extended the international design (grade 4) with an additional grade 3 sample; the average age differs by 12 months.

A closer look on table 3 reveals considerable differences between the countries' average age but also suggests some stability within countries. Taking different grades into account, within country differences are rather small. In Italy and the USA the average age for the 4th graders ranged between 116 to 118 and 120 to 124 months respectively in RL 1991 and the following studies.

Table 3: Age in month by study and country

Study	Hungary		Italy		Sweden		USA	
	Age Mean/SD	Grade	Age Mean/SD	Grade	Age Mean/SD	Grade	Age Mean/SD	Grade
SSS 1970	127/3	4/5	127/4	4/5*	125/4	3/4	127/5	4/5*
RL 1991	112/7	3	118/5	4	117/4	3	120/7	4
RL 2001	116/7	3	118/4	4	117/4	3	120/8	4
PIRLS 2001	128/6	4	118/4	4	118/4 130/4	3 4	124/5	4
PIRLS 2006	128/6	4	116/4	4	130/4	4	121/6	4

* A few students were also selected in from other grades (see table 1)

Average age in SSS has not been discussed so far, since the sampling strategy was different in comparison to the other studies. As mentioned above, SSS applied an age-based sampling, i.e. students

from different grades participate in the study. Table 4 contains information on the average age per grade in the 1970s study. It is quite obvious that the students in the sub-samples do not represent the whole age range in their grades. Mean age differences between two grades is not 12 but about 6 months in Hungary, Sweden and the USA. Such a difference underline that the sample of the upper grades contain rather young students, whereas the lower grade consists of rather old students. In Italy the age difference between grade 4 and 5 is just 2 months, but as the sample is very small (n=159) it might represent a very specific group of students.

Table 4: Age in months by country and grade in SSS

Grade	Hungary	Italy	Sweden	USA
	Age Mean/SD	Age Mean/SD	Age Mean/SD	Age Mean/SD
3	-	-	122/3	-
4	125/3	125/4	128/2	124/3
5	131/1	127/4	-	129/4

Significance of age and schooling on the comparability

This section briefly summarizes the research on the effect of schooling and age on reading literacy. This is important as the previous section on the amount of schooling and age pointed out differences between studies and countries in this respect. Such differences affect trend analyses if schooling and age have an impact on students' reading literacy. Consequently, a change in reading scores might not reflect actual change in reading literacy but differences in the samples.

The effects of both, schooling and age on general abilities (e.g. Cahan & Cohen, 1989; Ceci, 1991; Cliffordson & Gustafsson, 2008) as well as knowledge and skills in the school context (e.g. Cliffordson, 2008; Luyten, 2006) have been examined and proved in several studies. Thus, researchers controlled for such differences in their studies. In order to adjust the country scores, Rindermann (2007) estimated an effect of 42 points of an extra year of schooling on students' skills and abilities (on a scale with a mean of 500 and a standard deviation of 100). He took the results from several large-scale assessment studies on intelligence and subject specific competences and computed the mean differences between different grades, i.e. the adjustment of 42 point is the average gain in intelligence, mathematics and science over one year. It should be noted that such a global measure hides variations in the amount of gains. The effect was comparatively small for intelligence (20 points) but about two times larger for mathematics and science. Also, in lower grades the difference between two adjacent grades was greater than in higher grades.

The trend study focuses on students' reading literacy at the end of primary school. Referring to this,

the 2001s and 2006s PIRLS study enable analyses on the development of reading literacy in one year, since Iceland, Norway and Sweden expanded the international design that targets on grade 4 with additional samples from adjacent grades. In PIRLS 2001, Swedish 4th grade students outperformed their 3rd grade fellows by 41 points (500/100 scale; Mullis, et al., 2003). In 2006, a similar overall effect was found in Iceland and Norway. The differences between 4th and 5th grade students were 39 and 43 points respectively (Mullis, et al., 2007). Students from different grades differ in age as well as in the amount of schooling they received. In order to decompose the effect of age and schooling, Van Damme, Vanhee and Pustjens (2008) applied regression-discontinuity analyses on the PIRLS data. They found a net effect of an extra year of schooling of 10 (Iceland), 12 (Norway) and 26 (Sweden) points; with respect to age and the effect of one year was 28 (Iceland), 30 (Norway) and 16 points (Sweden).

Luyten (2006) analyzed TIMSS-95 data from 8 countries on students' achievement in mathematics and science at the end of primary school and decomposed the effect of aging and schooling. In mathematics, the analysis revealed that on average 60 percent (science = 53 percent) of the difference between students from two adjacent grades are due to schooling and 40 percent (science = 47 percent) due to aging.

A first tentative conclusion is that the observed differences in the amount of schooling and age should be considered in trend analyses. In the following section we will present strategies to control for such differences and discuss if they are applicable for our study.

Strategies to adjust for differences in age and schooling

Comparison derives from the assumption of having groups, which are comparable in congruent studies. However, as shown above, the sub-samples in this study are not comparable with respect to schooling and age. In this regard, we will present different approaches of how to deal with such differences in the following section.

In PIRLS 2001 and 2006 Italian students were almost one year younger than their fellows from the other countries although all students were in grade 4. For the purpose of studying trends across countries, this is not necessarily a problem since trend analyses focus particularly on change or development over time and not on a particular achievement level. Therefore, adjustments for age differences need not to be accounted for between countries but within countries, i.e. the aim is not to receive comparable age samples across countries but within countries over time. This approach is also called the differences-in-differences method (c.f. Angrist & Pischke, 2009), and the idea behind this strategy is simple: The countries might be observationally different (e.g. in age), but, as long as this observational difference is constant over time, it can be "differenced out". According to this consideration we will limit the following argumentation to strategies that enhance comparability of

differences within countries over time.

Data

The study on hand makes use of reading literacy data from all previous IEA reading literacy studies, namely the Six Subject Survey (1970), the Reading Literacy Study (1991) and its repeat (2001), PIRLS 2001 and PIRLS 2006. The database contains test scores and background information from 86604 students. Links between the studies were established by overlapping tests and samples. These overlaps made it possible to estimate a common IRT-scale for all studies with the aid of the PARSCALE 4 software. Data from Hungary, Italy, Sweden and the USA, equally weighted, was used to estimate a two parameter IRT-model and student reading ability (Rosén & Strietholt, 2010a, 2010b). The scale has a mean of 500 and a standard deviation of 100. Note that this scale is not the same as used in PIRLS since the transformation of the scores is based on the mentioned four countries and five studies. The sampling weights are from the dataset, and in the analysis they were used to correct for different selection probabilities. The analyses were conducted with Stata 10 and took cluster sampling into account.

Strategy 1: Focus on sub-samples

A simple and straightforward strategy to control for differences in age and grade is to focus on students. With respect to grade this is straightforward: In 2001 data from two grades is available for Hungary and Sweden. The difference between grade 3 and 4 students is 36 points in Hungary and 52 points in Sweden. Table 5 also contains information on the differences in SSS. Here it is important to note that the students were not sampled at random in each grade: students from lower grades are relatively old whereas those from higher grades are relatively young.

Table 5: Grade differences in reading achievement scores

	SSS 1970				**PIRLS/RL 2001**	
Grade	**Hungary**	**Italy**	**Sweden**	**USA**	**Hungary**	**Sweden**
3	-	-	521	-	466	477
4	471	485	567	472	502	529
5	492	552	-	511	-	-
Difference	21	67	46	39	36	52

In order to control for such age differences a similar strategy can be applied: Students of the same age might be more appropriate for comparisons. To evaluate this strategy, grade 4 students from the USA of the 1970s and 1991s sample qualify for this purpose. For 10 months test scores from a sufficient number of students (n<50) are available for such comparisons between 1970 and 1991. Table 6

summarizes the mean test scores of students at a different age as well as the mean difference between the students at the same age in both studies. On average, the students from 1991 outperform those from 1970 by 33 points. The largest differences are for 128-months-old students (47 points), the smallest for the 129-months-old students (5 points). Table 6 (at the bottom) also shows the mean scores of all students from grade 4 in the respective studies. On average, the students from 1991 achieve by 38 points higher test scores than their fellows from 1971. As described in table 3 and 4, the average age was 124 months in SSS and 120 months in RL. Note that the sample sizes are considerable larger since no student had to be excluded for this comparison.

It seems reasonable that the test scores from a sample with relatively young students are biased downwards. Focusing on sub-samples of students at the same age should compensate such a bias. The results of the analysis presented above suggest that age differences really have an impact on the mean achievement – but in a way that is not in line with the other research. The analysis revealed a reverse effect and suggests adjusting the mean score downwards by 1.25 points (5 points for 4 month) when average age increases by 1 month. However, even if the results do not seem to be plausible, a difference of 1.25 points per month is rather small and do not affect the results substantially.

A crucial fact that is connected with the strategy to focus on sub-samples is the loss of power. The constraint to focus on specific groups of students is attended by a reduction of the effective sample size and larger standard errors. In the above sample, only 2546 of 6443 students from the RL 1991 data were used since the overlap in age to the SSS students was limited to a few months. We also applied similar analyses to other countries, but it turned out that only smaller comparable age groups were available in other samples from the SSS. Against this background, the efficiency of this approach might be questioned.

Table 6: Age differences in reading achievement scores

Age in month	SSS 1970			RL 1991			Difference between SSS and RL
	Mean	S.E.	N	Mean	S.E.	N	Mean Difference
121	486	6.2	261	525	3.3	781	39
122	492	6.3	222	529	5.3	319	37
123	489	6.5	237	525	5.5	311	36
124	483	7.6	165	515	5.6	278	32
125	476	7.9	147	501	6.8	186	25
126	439	9.3	110	474	7.0	188	35
127	435	11.8	59	459	7.8	156	23
128	408	9.5	61	455	8.5	122	47
129	452	12.3	55	461	9.7	91	9

130	409	11.0	54	437	7.8	114	29
Mean test score of students at age 121 to 130 month	**472**	**-**	**1371**	**505**	**-**	**2546**	**33**
Mean test score of students at any age	**472**	**4.2**	**1675**	**509**	**2.9**	**6443**	**38**

Strategy 2: Correction model

Another strategy is to set up a correction model to adjust the initial scores. Equation (1) presents such an adjustment model, where Y is the value to adjust for. U, V and W are differences in age, grade and test month that vary between the samples. The lowercase letters a, b and c are constants and describe by how many points the initial scores have to be corrected for when U, V and W change by one unit.

(1) $Y = aU + bV + cW$

The crucial point is to find appropriate values for a, b and c. Some researchers have conducted analyses on IEA data in order to estimate the effects of age and grade on student achievement. But, as mentioned above, most research has been done on mathematics and science and, additionally, the results vary considerable between different grades. We therefore suggest a focus on research on students' reading literacy at the end of primary school. However, the amount of research on this specific issue is limited since data from two adjacent years is only available from the Icelandic, Norwegian and Swedish extensions of PIRLS. In all three countries the development of reading literacy in one year was about 40 points. In Sweden roughly 2/3 was due to schooling and 1/3 due to aging whereas the proportion was reverse in Iceland and Norway (see above and Van Damme, et al., 2008). Such between-country-differences were also observed in Luytens (2006) study, which contained eight countries, but on average the impact of schooling was somewhat larger than the impact of age (60 vs. 40 percent in mathematics and 53 vs. 47 in science).

The above-mentioned studies did not address the question of effects of different test months. However, it seems reasonable to assume that the development of students reading literacy in one school year is approximately linear. Therefore, to get an estimate for what students learn in one month the gain in reading achievement in one school year can be divided by 12. This assumption of linearity might appear somewhat oversimplified because of school vocation, but we do not think that this affects the estimate substantially.

For above reasons, we place the following values into equation (1) and get a correction formula for age, grade and test month adjustment:

(2) $Y = \text{age}*23 + \text{grade}*23 + \text{test month}*2$

The choices for correction factors were based on the following considerations. Students from higher grades outperformed their younger fellows by 41, 39 and 43 points (on the PIRLS scale; Mullis, et al., 2003; Mullis, et al., 2007) in Iceland, Norway and Sweden. In our trend study we linked Hungarian and Swedish data from grade 3 (RL 2001) and 4 (PIRLS 2001) and found a grade difference of 36 and 52 points (on the trend scale). We used the Swedish measures to transform the Icelandic and Norwegian ones on the trend scale. In a second step we averaged over the four country means and yield a value of 46 points; this equals what students learn over one year. In order to distinguish between the age and the grade effect, we applied a similar strategy and averaged the values reported in Van Damme, Vanhee and Pustjens (2008). On average, 39 percent of what students learn in one year is due to schooling and 61 percent due to aging. This contrasts with the findings of Luyten (2006) for mathematics and science. Particularly, the findings from Norway are remarkable since information on reading, mathematics and science are available and can be compared: In reading 71 percent is due to aging (29 percent to schooling) but in mathematics and science just 40 respectively 47 percent (60 respectively 53 to schooling). For that reason, we assume that the findings on reading are somewhat biased and adopt an equal factor for age and schooling as correction factors. Therefore, we split the overall effect of what students learn over one year (46 points) into 23 points for the age effect and 23 points for the schooling effect. For differences in the test months we consider 2 points per month (= 23/12).

Table 7: Raw scores and adjusted scores for the correction model

SSS 1970						
Grade		**Age difference**	**Test month difference**	**Initial unadjusted score**	**Adjust-ment**	**Adjusted score**
3	Sweden	+4.4	not recoded	521	-9	512
4	Hungary	-2.5		471	5	476
4	Italy	+6.5		485	-13	472
4	Sweden	-1.9		567	4	571
4	USA	+2.3		472	-5	467
RL 1991						
3	Hungary	-3.9	not recoded	449	8	457
3	Sweden	+1.0		509	2	511
4	Italy	+0.1		486	0	486
4	USA	-2.1		509	4	514
RL 2001/PIRLS 2001						
3	Hungary	baseline	baseline	466	-	-
3	Italy			477	-	-
4	Hungary			502	-	-
4	Italy			498	-	-
4	Sweden			529	-	-
4	USA			498	-	-
PIRLS 2006						
4	Hungary	-0.14	-1.0	505	2	508
4	Italy	-1.64	-2.0	506	7	513

4	Sweden	+0.26	+0.3	504	-1	502
4	USA	-0.93	-2.1	490	6	496

We applied equation (2) to compute adjusted means for the 3rd and 4th grade samples. As a starting point we used 2001 and adjusted the samples from the other years in such a way that they are comparable to the 2001s sample. These corrections take differences in the test months as well as in age into account. Table 7 shows unadjusted and adjusted scores for the respective studies and countries. As we used 2001 as a baseline, no adjustments have been made for this study. Note that the corrections do not adjust for different age and test months between countries. As mentioned before, we did not aim to do such corrections since they would not affect the measures of development. However, the largest adjustments have been made for the Italian SSS sample (13 points), the others are all below 10 points.

Strategy 3: Variation in test month within countries and studies

A third strategy to control for differences in the test months is to use variations in the test months itself. Since data was collected in different months in some samples (see table 2) such a variation might be used to estimate the effect of test months. The samples from the USA in PIRLS 2001 and 2006 as well as the Swedish PIRLS 2006 contain test scores from a sufficient number of students from more than one month. However, preliminary analyses revealed that such information could not be used for the estimation of the effect of the test month. Table 8 lists the average test scores for students that were tested in different months. The results for the USA are odd for 2001 since the students tested in April receive lower scores than the students tested in May. In 2006, an increase in the test scores from January to April can be observed, whereas there is a decrease in May and June. In Sweden the data was mainly collected in April and May. Students who were tested in May outperform those tested in April by 7 points.

In summary, from a theoretical point of view, the test month might affect the results. Differences in the test months influence the idea of comparable groups between countries and years. However, empirically, the analysis of the limited data on this issue yields inconclusive results. Therefore, we did not further apply the 3rd strategy in the data analysis.

Table 8: Test scores by test month

	USA 2001		USA 2006		Sweden 2006	
	Mean score	N	Mean score	N	Mean score	N
January	-	-	481	1504	-	-
February	-	-	493	1620	-	-
March	-	-	496	758	522	61
April	504	2085	504	483	501	2765
May	494	3501	486	684	508	1568

June	-	-	476	140	-	-

Conclusion

In the present study, we have identified differences in age and schooling of the four countries in the samples that were includes in the trend analysis on reading achievement from 1970 to 2006. Such differences have to be considered since age as well as schooling is related to the development of students' achievement. We presented miscellaneous strategies to adjust the mean test scores in order to enhance comparability. Besides, we applied the strategies to the data from the trend study to evaluate if they are suitable and by how far corrections affect the results. In reverse, the amount of corrections also gives an impression of the degree to which unadjusted scores are trustworthy?

According to the analyses, a correction model seems to be the most appropriate approach to control for differences in age and schooling. The advantage of such a model is that it is applicable to all samples if information on age, test month and grade is available. However, the crucial point is to find proper values for the correction equation, which is not trivial. Information on the effect of age and schooling on reading literacy at the end of primary school is very limited. Additionally, considerable variations between different studies on the effect size of age and schooling raises the questions if general values are suitable for all countries and studies in the trend study. Therefore, further research has to be carried out in order to reach a convincing answer.

By how far do differences in age and schooling restrict comparability of the respective samples in the trend study? Our analyses provide a first insight into this question. The adjustments we made are on average 5 points; only one adjustment exceeds 10 points.

In summary, the results from our study revealed that the comparability of the samples from the respective studies and countries could be considered as a function of age and schooling. Differences in these factors have a moderate effect on the actual comparability. Even if such differences would not hide substantial developments in reading literacy it seems advisable to adjust the initial scores. Our correction model provides an approach for such adjustments. Further research has do be done in order to evaluate the robustness of the model.

References

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics. An empirist's companion*. Princeton.

Cahan, S., & Cohen, N. (1989). Age versus Schooling Effects on Intelligence Development *Child Development, 60*(5), 1239-1249.

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology, 27*(5), 703-722.

Cliffordson, C. (2008). *Effects of Schooling and Age on Performance in Mathematics and Science: A Between-Grade Regression Discontinuity Design Applied to Swedish TIMSS 95 Data*. Paper presented at the 3rd IEA International Research Conference (IRC 2008), Taipei City, Taiwan.

Cliffordson, C., & Gustafsson, J.-E. (2008). Effects of Age and Schooling on Intellectual Performance: Estimates Obtained from Analysis of Continuous Variation in Age and Length of Schooling. *Intelligence, 36*(2), 143-152.

Elley, W. B. (1992). *How in the World do Students Read? IEA Study of Reading Literacy*. Hamburg: Grindeldruck.

Elley, W. B. (1994). *The IEA Study of Reading Literacy: Achievement and Instruction in Thirty-two School Systems*. Oxford: Pergamon Press.

Hansson, G. (1975). *Svensk skola i internationell beslysning II [The Swedish school in an international setting II]*. Stockholm: Almqvist & Wiksell.

Luyten, H. (2006). An Empirical Assessment of the Absolute Effect of Schooling: Regression-Discontinuity Applied to TIMSS-95 *Oxford Review of Education, 32*(3), 397-429.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., & Kennedy, A. M. (2003). *Trends in Children's Reading Literacy Achievement 1991–2001: IEA's Repeat in Nine Countries of the 1991 Reading Literacy Study.* Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Gonzales, E. J., & Foy, P. (2003). *PIRLS 2001 International Report: IEA's Study of Reading Literacy Achievement in Primary Schools in 35 Countries*. Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 International Report. IEA's Progress in International Reading Literacy Study in Primary School in 40 countries.* Chestnut Hill, MA: Boston College.

Rindermann, H. (2007). The *g*-Factor of International Cognitive Ability Comparisons: The Homogeneity of Results in PISA, TIMSS, PIRLS and IQ-Tests Across Nations *European Journal of Personality, 21*(5), 667-706.

Rosén, M. (2006). *Analysing trends in levels of reading literacy between 1970 and 2001 in Sweden.* Paper presented at the 2nd IEA International Research Conference, Washington.

Rosén, M., & Strietholt, R. (2010a). *Choosing between the 1-, 2- and 3-PL Models in a trend study.* Paper presented at the symposium "Modelling Longitudinal Data" at the ECER 2010 "Educational and Cultural Change" in Helsinki, Finland.

Rosén, M., & Strietholt, R. (2010b). *Linking reading literacy tests for a 35 year trend study. Analysis*

of the bridge items. Paper presented at the symposium "Modelling Longitudinal Data" at the ECER 2010 "Educational and Cultural Change" in Helsinki, Finland.

Rosén, M., & Strietholt, R. (2010c). *Trends in reading literacy from 1970 to 2006. A comparison on 9-10 year olds in Sweden, Hungary, Italy and the USA.* Paper presented at the 4th IEA Research Conference, Göteburg and Oslo.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for General Causal Inference*. Boston, MA: Houghton Mifflin.

Thorndike, R. L. (1973). *Reading Comprehension Education in Fifteen Countries: An Empirical Study*. Stockholm Almquist & Wiksell.

Van Damme, J., Vanhee, L., & Pustjens, H. (2008). *Explaining reading achievement in PIRLS by age and SES*. Paper presented at the The 3rd IEA International Research Conference, Teipeh, Taiwan.