Maciej Jakubowski
The Robert Schuman Centre for Advanced Studies, European University Institute, Italy
Department of Economics, Warsaw University, Poland

Effects of tracking on achievement growth

Exploring difference-in-differences approach to PIRLS, TIMSS and PISA data

1. Introduction

This paper discusses difference-in-differences (DD) methodology employed to assess tracking effects using data from international educational surveys of 4th graders (PIRLS 2001, TIMSS 2003) and 15-year-olds (PISA 2000 and 2003). We define tracking system as the one where at some point students are separated into schools which differ in educational programme or objectives. Thus, this is institutional tracking which in most countries happens in secondary education but at different age. The fact that none of the countries separate students into different programmes earlier than the 4th grade makes possible examining tracking effects using difference-in-differences approach by comparing differences in achievement between primary and secondary school students among tracking and non-tracking countries, namely between countries which track 15-year-old students and countries which do it later. The main reference point of this paper is the seminal work by Eric Hanushek and Ludger Woessmann who assessed tracking by implementing DD method to analyze data from PIRLS, PISA and TIMSS. This paper builds on this work providing robustness checks of their results and extends methodology analyzing student-level micro data.

Hanushek and Woessmann claimed that there is no gain in mean performance from tracking and that tracking increases educational inequalities. Specifically, they showed that standard deviation of scores in secondary schools in tracking countries is higher than in non-tracking countries controlling for scores variation in primary schools in those countries. We checked in several ways whether these results are robust. Using individual data we restricted the samples from different surveys to make them comparable and additionally adjusted for differences in average students age using regression framework. We found evidence on the negative impact of tracking on mean performance, however, it was emphasized that results could be biased by systematic differences between tracking and non-tracking countries in other dimensions than tracking policy. Arguments were provided why straightforward comparisons of score variation measured by different surveys is not valid. Several adjustments to make such comparisons more legitimate were discussed and estimates of tracking effects for different specifications were provided. Results suggest that there is no clear evidence that tracking has negative impact on inequalities.

The final section of the paper contains analysis of student-level. Using simple regression framework we checked whether treatment effects are homogenous. The presumption was that tracking should affect more negatively students with less advantageous family background while there should be no impact on students with more advantageous background. The argument was that while the former group is usually tracked into vocational schools the latter group stay in comprehensive school system similar to that in the non-tracking country. However, we found no evidence suggesting that tracking affects more heavily students with less advantageous background. It seems that there are systematic differences in achievement growth in tracking and non-tracking countries considered which homogenously affects all students. Thus, we believe that there are other causes than tracking of lower performance in the group of tracking countries. In other words, evidence provided in the paper suggests that earlier results were confounded with other factors than tracking.

The paper is organized as follows. Section 2 discuses methodology. Section 3 describes data used in the study and the set of countries considered. Section 4 provides analysis within the difference-in-differences framework developed by Hanushek and Woesmann. The main purpose here is to check whether their results are robust to differences in sample design and methodology of international educational surveys. Section 5 describes analysis done with slightly different DD method and uses individual student-level data to assess heterogeneity of tracking effects. Section 6 concludes.

2. Methodology

We define framework using the approach used in treatment evaluation literature (Lee, 2005). Assume that we have individual level cross-sectional data from two periods: t=0 and t=1 (in our case from primary and secondary school, respectively). Tracking is the 'treatment' imposed in period t=1. We assume that outcome y_t measures the same domain of academic achievement in both periods. For each individual we observe $\boldsymbol{y_0}$ or $\boldsymbol{y_1}$ but not both, because only repeated cross-sections are available. Let d be the treatment indicator. Then d=1 when student is in the tracking system and d=0 otherwise. Let c=1 if student is in the tracking country and c=0 otherwise. There are eight potential outcomes. Some of them are observed in reality and others are not. Those not observed are called counterfactuals and have to be estimated from the observed outcomes of matched individuals.

Using this notation the usual DD estimator is:

$$DD = \{ E[y_1|d=1,c=1] - E[y_0|d=0,c=1] \} - \{ E[y_1|d=0,c=0] - E[y_0|d=0,c=0] \} \quad [1]$$

With several assumptions DD identifies average treatment effect on the treated (ATT), which in this case is the average effect of tracking on students in tracking countries (see Lee, Kang, 2005; Lee, 2005). In the case of panel data DD is estimated by simple comparison of average of individual differences between t=0 and t=1 in the country (group) with and without treatment. However, if only cross-sectional data are available then we observe only one outcome per individual and the problem of whether outcomes of different individuals are comparable arise. For example, if we observe $y_1(d=1)|c=1$ for *i*-th individual then we do not observe $y_0(d=1)|c=1$ for this individual. However, we could observe this outcome for similar *j*-th individual at t=0. It is assumed that if cross-sectional data are representative for the population of interest then the average outcome in the sample at time t=1 can be compared with the average outcome in the sample at time t=0 both in treated and non-treated countries. This is based on assumption that in the fully representative samples individual characteristics are perfectly balanced. Thus, we calculate average outcomes from four representative samples needed to calculate DD from equation [1].

To be more specific consider the case of PIRLS and PISA. From PIRLS data the average achievement in primary school (t=0) can be calculated for any participating country which should be representative for the population of 4th graders. Doing the same for PISA data one can obtain estimate of average achievement in the population of 15-year-olds. However, some countries participating in PISA have already tracked students by separating them into different types of schools, usually comprehensive and some type of vocational. Calculating average achievement for tracking and non-tracking countries in PIRLS and PISA we obtain four outcomes which can be plugged into equation [1] to obtain DD estimate of tracking effects:

$$DD = [E(Y_{PISA}) - E(Y_{PIRLS})|\textit{tracking-countries}] - [E(Y_{PISA}) - E(Y_{PIRLS})|\textit{non-tracking countries}]$$

where Y is the average achievement in a country measured through specific educational survey.

The main benefit from DD approach should be clear now. We compare changes within countries which limits the bias caused by unobserved or not controlled differences between countries. Simply, comparing achievement in tracking and non-tracking countries is not a satisfactory approach to assess effects of tracking. Countries achievement levels are affected more heavily by other policies than tracking and are driven by other than educational differences between them. Using DD approach to estimate tracking effects we don't have to care about between countries differences in early achievement or other stable features. However, we still have to consider other characteristics which affect achievement growth or makes outcomes in two periods incomparable.

In practice DD is estimated using regression analysis. Let *t* be the dummy indicator of time, *d* dummy indicator of treatment (tracking) and *dt* dummy indicator of treated units (secondary school students in tracking countries, in fact, this is interaction term of time and treatment). Then DD estimator of tracking can be obtained by estimating following equation:

$$Y = \alpha_0 + \alpha_1 t + \alpha_2 d + \beta dt + \varepsilon \quad [2]$$

where β is the DD estimate of interest. We will call this nonparametric DD estimate and if the observation is any country-level statistic then we will call it nonparametric country-level DD estimate.

In the paper which is the main reference point for this study Hanushek and Woessmann used quite different specification assuming linear relation between outcomes from *t*=0 and *t*=1. They estimated following equation using country-level data:

$$Y_1 = \alpha_0 + \alpha_1 Y_0 + \beta d + \varepsilon \qquad [3]$$

where β is believed to be the DD estimate of interest, and Y_1 and Y_0 are outcomes in t=1 and t=0 respectively (e.g., average achievement in PISA and average achievement in PIRLS). Thus, DD here is the difference in intercept between tracking and non-tracking countries in the regression equation where early achievement is used to explain secondary school achievement. We will call this country-level parametric DD. In theory it is possible to add additional explanatory variables which are supposed to affect achievement independently from tracking. In practice the small number of countries considered importantly limits this option. Equations [2] and [3] can be used to estimate DD effects on any chosen statistic, e.g. standard deviation or median.

There are two crucial assumptions in the DD approaches presented above. We will discussed them not in general but in relation to characteristics of PIRLS, TIMSS and PISA (for general discussion of DD approach to cross-sectional and other types of data see: Lee, Kang, 2005; Meyer, 1995). First of all, it is assumed that samples of students collected at *t*=0 and *t*=1 are fully comparable and representative to similar populations of students which differ only by the fact that one is in primary and other is in secondary school. In other words, it is assumed that students sampled in PIRLS have on average similar characteristics to those sampled in PISA. As we will show this is in fact not true because of differences in study design between PIRLS (or TIMSS) and PISA, mainly because in PIRLS the population of interest is defined by grade while in PISA the age criterion is crucial.

The second assumption is usually called "same time effect". To identify treatment effect it is needed that baseline response of those treated would be the same as in the control group (untreated) if their would be not treated. More specifically it is assumed that:

$$\mathrm{E}[(y_1 - y_0)|d=0, c=1] = \mathrm{E}[(y_1 - y_0)|d=0,c=0] \qquad [4]$$

In the case of PIRLS and PISA this means that achievement would change by the same magnitude in tracking and non-tracking countries if there will be no tracking at all. It is easy to imagine that there are common characteristics which influence achievement growth in tracking countries differently than in non-tracking countries. One should control for such characteristics to obtain non biased estimates. While $y_1(d=0)|c=1$ is not observable it is not possible to fully test this assumption, but one can assess whether effects of interest have similar and reasonable impact on different groups. Especially if in the treated countries there is a group which is believed to be less or more heavily affected by treatment. In this paper we do this by looking at heterogeneity of tracking effects on student groups defined by family background. We assumed that tracking should have negligible impact on students with the most favourable background and more substantial impact on students with the less advantageous background. Family background is measured here by parental education (PIRLS and PISA) and the number of books at home (TIMSS and PISA). Such study design is sometimes called difference-in-differences-in-differences (DDD) and could be implemented within the regression framework which opens the possibility to additionally control for individual characteristics (see Gruber, 1994). Let $\mathbf{x}_{\mathrm{isc}}$ be the vector of individual, school or country characteristics we want to control for, g_{isc} be a dummy variable indicating group which we believe is affected by treatment, and *i*, *s*, *c* indexes individuals, schools (classes) and countries, respectively, then the general form could be written as follows:

$$y_{isc} = \alpha_0 + \alpha_1 t_{isc} + \alpha_2 d_c + \mathbf{x}'_{\mathbf{isc}}\boldsymbol{\alpha}_3 + \beta_1 dt_{isc} + \beta_2 gt_{isc} + \beta_3 dg_{isc} + \gamma dtg_{isc} + \varepsilon_{isc} \qquad [5]$$

where for example *dt* or *dtg* are interaction terms of *d*, *t*, and *d*, *t*, *g*, respectively. Now, the parameter γ is of interest and estimates tracking effect on the group indicated by *g*. Interaction terms of treatment and time with group indicator or other characteristics help us to exclude the possibility of different outcome trajectories for students with observable characteristics.

3. Data

This study uses international datasets provided by organizers of PIRLS, TIMSS and PISA. Data as well as documentation are accessible online: for PIRLS and TIMSS see http://timss.bc.edu, for PISA see www.pisa.oecd.org and links available there. All estimates used in this paper were independently calculated from these datasets and checked with original reports from survey organizers. They slightly disagree because the sample of countries is different. In this research only countries which took part in PIRLS or TIMSS and PISA are analyzed. Additionally, data from some countries were not used in official reports, for example the Netherlands in PISA 2000 or UK in PISA 2003, but we decided to analyze them. Moreover, we separated data for England and Scotland using indicators given in the datasets and treat them as two distinct countries recognizing the fact that these are two separate school systems and surveys were organized independently.. We also used only Flemish Community data for Belgium because only those are available in TIMSS[1].

The Table below contains names of countries considered in three comparisons studied here and indicates which country was considered as tracking or non-tracking. Tracking dummy was created based on the data from Eurydice (see www.eurydice.org) or in the case of non European countries from national websites.

Table 1. Countries participating in surveys and considered in the research.

PIRLS and PISA 2000	PIRLS and PISA 2003	TIMSS 2003 and PISA 2003
Non-tracking countries:	*Non-tracking countries:*	*Non-tracking countries:*
Argentina	Canada	Australia
Canada	England	England
England	Hong Kong	Hong Kong
Hong Kong	Iceland	Japan
Iceland	Latvia	Latvia
Israel	New Zealand	New Zealand
Latvia	Norway	Norway
New Zealand	Scotland	Scotland
Norway	Sweden	Tunisia
Scotland	Turkey	United States
Sweden	United States	
United States		
Tracking countries:	*Tracking countries:*	*Tracking countries:*
Bulgaria	Czech Republic	Belgium (Flemish Community)
Czech Republic	France	Hungary
France	Germany	Italy
Germany	Greece	Netherlands
Greece	Hungary	Russian Federation
Hungary	Italy	
Italy	Netherlands	
Macedonia	Russian Federation	
Netherlands	Slovak Republic	
Romania		
Russian Federation		
23 countries, 218013 observations	20 countries, 210693 observations	15 countries, 152777 observations

Average achievement scores are given in the appendix in the Table A1, separately for each country and survey. It have to be emphasized that all achievement scores were standardized to have mean 500 and standard deviation 100. This means that mean performance of countries and its variation differ from those

[1] While there is no indicator which can be used to separate data from different communities in PISA there are differences in questions which were used to identify Flemish Community schools.

published in official reports where similar standardization was done for all countries participating in a particular survey, which was about twice the number of countries investigated here. However, standardization makes results from different surveys comparable in the group of countries investigated in this research. Regardless these transformations ranking of countries according to their mean performance was preserved which means that it was roughly the same as published in official PIRLS, TIMSS and PISA reports.

Part of the results presented in the paper where obtained from the sample limited to the chosen set of observations. Excluded observations could affect score distribution and final results. Thus, standardization was also done for this sample in the same way as for the whole sample. We do not present statistics for this sample because they are virtually the same as for the whole sample.

Organizers of PIRLS, TIMSS and PISA use similar methodology to produce achievement scores for all students. Details are given in the technical documentation available on surveys websites. The most important fact is that achievement scores are available as a set of five plausible values for each individual. Ideally one should repeat any analysis with five plausible values and final statistic of interest should be calculated as a mean of those five estimates. We followed this strategy only in the case of variance estimation were it could really affect results. In the analysis of average performance we used the first plausible value provided in each dataset. Examples given in surveys documentation as well as several checks done by us show that working with different plausible value did not change results in the case of mean performance and related regression analysis. It have to be said that this way our estimates of standard deviation are slightly smaller because we did not consider variation produced by the process of plausible values imputation or test measurement. Nevertheless, it surely did not affect conclusions we made in this study.

In some sections of the paper additional variables are used which in most cases were constructed from original data to obtain the same definitions. Dissimilar definitions and characteristics collected in distinct surveys narrow the set of potential covariates that could be used to analyze pooled data. Modest set of social and economic characteristics collected in PIRLS and TIMSS is importantly limiting any research of the kind proposed here. Nevertheless, it is still possible to construct variables which are similarly defined in all surveys. In our case it was gender, parental education, number of books at home, and indicators of whether students was born in the country of the test and does she or he speaks the language of the test at home. Description of these variables separately for different surveys, tracking and non-tracking countries are presented in the Table A2 in the appendix.

4. Checking the robustness of country-level difference-in-differences approach

In their seminal paper Eric Hanushek and Ludger Woessmann proposed the difference-in-differences approach to assess the effects of tracking on achievement growth and distribution. They used country-level statistics (mean, standard deviation, differences between percentiles) to estimate simple regressions where PISA or TIMSS 8th graders results where used as an independent variable and PIRLS or TIMSS 4th graders results as a proxy of early achievement and independent variable together with a tracking dummy equal 1 for countries where secondary school students are separated between schools with different educational programmes. Description of this approach was given in the section 2.

In this section we follow HW methodology analyzing data from PIRLS, TIMSS and PISA, to check robustness of their results. We began with estimates obtained for full sample but with different outcome measures and then put some restrictions to make data from different surveys more comparable using the possibilities open by using micro student-level data. We started with analysis of tracking effects on the achievement growth. First, we estimated tracking effects for different, more specific achievement measures in reading, mathematics, science and problem solving. Second, we considered differences in average age of 4th graders samples in PIRLS and TIMSS and the sample in PISA. This was done by adding to the regression equation variable equal to the difference between average age of students in each country taking one of the tests. Third, we limited the analysis to the sample of native students. Finally, we put all the restrictions altogether to compare with results obtained for the full sample. In what follows we call the dataset with more comparable data from different surveys the restricted sample.

The second part of this section concentrates on educational inequalities. Using approach similar to that of Hanushek and Woesmann it was analyzed how tracking affects the change in achievement scores dispersion in countries considered. Again, we started with complete dataset to compare results with those obtained from the restricted sample. Restrictions affected results because countries' score distributions were strongly affected by the way they sampled students. Moreover, we argue that results obtained from the restricted dataset are more valid because samples from PIRLS, TIMSS and PISA were made as similar as possible.

In the last section analysis is presented where tracking effects were separately estimated for groups of students with different family background. We used two simple indicators of family educational resources: the highest level of parental education (in the case of PIRLS and PISA) and the number of books at home (in the case of TIMSS and PISA). This way we wanted to check whether tracking effects are homogenous or interact with family background.

Main results are presented in the text while additional Tables are presented in the Appendix.

4.1. Effects of tracking on achievement in reading, mathematics, science and problem solving.

4.1.1. Effects of tracking on overall achievement based on full sample.

We began with analysis very similar to that of Hanushek and Woessmann. From individual data average overall achievement in several subjects was estimated using weights given in the datasets. This way we obtained data comparable to those analyzed by Hanushek and Woessmann, but standardized at the individual level in the sample of analyzed countries. That differs from their approach where standardization was made with country-level averages. In addition, we compared all domains tested in PISA: reading, mathematics, science and problem solving relating them to similar scores from PIRLS and TIMSS (problem solving in PISA was related to science and mathematics in TIMSS). The results are presented in the Table 2.

Table 2. Country-level DD in overall mean achievement in reading and in reading blocks

	PIRLS and PISA		**TIMSS 2003 and PISA 2003**			
	2001/2000 Reading	*2001/2003 Reading*	*Mathematics*	*Science*	*Problem solving and mathematics*	*Problem solving and science*
Tracking	-35.35*** (11.74)	-26.75*** (7.83)	-17.54 (17.04)	-8.01 (10.80)	-19.65 (16.96)	-11.67 (16.66)
4^{TH} grade achievement	0.66*** (0.14)	0.49*** (0.12)	0.66*** (0.13)	0.57*** (0.09)	0.72*** (0.13)	0.71*** (0.13)
Constant	183.59** (69.61)	261.18*** (60.47)	174.33** (66.05)	215.60*** (42.64)	144.59** (65.76)	145.41** (65.74)
Adj. R-squared	0.554	0.537	0.613	0.748	0.659	0.658
N	23	20	15	15	15	15

These results are qualitatively similar to those obtained by Hanushek and Woessmann. They suggest that tracking have strong negative impact on average reading literacy. Results for mathematics, science and problem solving are not significant but given the smaller sample and negative signs one could conclude that they support the claim that tracking has at least no positive effect on overall average achievement in these subjects.

4.1.2. Effects of tracking on different cognitive and content domains in reading, mathematics and science.

Typically in economic and sociological research of PIRLS, TIMSS and PISA only overall measures of subject specific achievement are employed. However, all international tests produce also subscales measuring achievement in several cognitive or content domains of one subject. For example, the PISA 2000 dataset contains plausible values of overall achievement in reading and at the same time three sets of plausible values in reading literacy subscales: retrieving information, interpreting texts, reflection and evaluation. Regardless that students were not tested in all domains the plausible values for every student were produced[2]. In this section subscales are used to check whether results are robust to differences in outcomes measured.

This is especially important in the DD approach proposed in this study where tests developed by different teams of researchers and based on distinct assumptions are compared. Any systematic differences between tracking and non-tracking countries in the content of subject-specific knowledge students are supposed to master during the lower and upper secondary education could heavily bias results. One could argue that tracking countries differently define their curricula for students in comprehensive and non-comprehensive schools while non-tracking countries have no possibility to do this. Thus, in the case of DD approach discrepancies between countries curriculum and psychological constructs measured in international tests could affect results

All same-subject subscales provided in PIRLS, TIMSS and PISA datasets were related to each other. The subscales created in PISA 2000 were already mention. Those were related to PIRLS subscales in reading: reading for literary experience, and reading to acquire and use information. In PISA 2003 we have four separate overall literacy scales for reading, science, mathematics and problem solving. However, given that mathematics were a main tested domain in 2003, it was possible to construct four mathematics literacy subscales: space and shape, uncertainty, change and relationship, and quantity. Those could be related to TIMSS 2003 subscales measuring content in mathematics: number, patterns and relationships, measurement, geometry, and data. Additionally, there are 3 subscales in cognitive domains in mathematics: applying, knowing and reasoning, as well as separate scores for life, physical, and earth sciences. PISA 2003 produced also scores in over-curriculum domain called problem solving, which was related here to overall and subscale scores in mathematics and science from TIMSS 2003.

Summing up, it was possible to compare several pairs of sub-domains tested in PIRLS, PISA and TIMSS:

- 6 in reading literacy for PIRLS/PISA 2000,
- 2 in reading for PIRLS/PISA 2003,
- 32 in mathematics for TIMSS 2003 and PISA 2003,
- 3 in science for TIMSS 2003 and PISA 2003
- 11 relating problem solving measured in PISA 2003 to all subscales in mathematics and science in TIMSS 2003.

It has to be said that this paper made no attempt to rethink the way distinct subscales should be related to each other. Instead, all possible pairs were compared to check whether estimated tracking effects are robust to differences in measured constructs and discrepancies in countries' curricula. The results are presented in the Appendix (see Tables A3-A6). All estimated effects of tracking on reading literacy where negative and highly significant and varied from -25 to -43 (from ¼ to almost ½ of standard deviation of scores). Among

[2] For details of how plausible values in subscales were constructed see: OECD 2002, 2005; Martin et al., 2003; 2004.

32 estimated effects of tracking in mathematics only 2 were positive. In science all tracking coefficients were negative as well as in problem solving measured in PISA 2003 and related to all subscales in science and mathematics in TIMSS 2003. In the case of TIMSS 2003 and PISA 2003 none of the estimated coefficients were significant at the 10% level, however, those estimates were obtained for the sample of 15 countries only. It is worth noting that in all regressions early achievement was highly correlated with achievement in secondary school regardless of the subject and subscale employed.

Great cohesion of these results suggest that negative effects of tracking are an artefact produced by differences in curriculum between tracking and non-tracking countries. However, it seems that the magnitude of tracking effect depends on construct measured through the test. The lowest estimates were nearly -½ of standard deviation but the highest were close to zero. Thus, changing outcomes measurement do not affect main conclusion that tracking has no positive effect on average achievement, but it clearly has an impact on tracking estimates magnitude.

4.1.3. Analysis restricted to native students

This section contains results of analysis done on the sample of native students only. One can assume that migrants should not be considered to assess tracking effects because some of them were not in the country of the test in the 4th grade or at least were not fully exposed to this country education system. Another argument is that if tracking effects do exist then they should also affect native students. Thus, all students who were born outside the country of the test were excluded from the sample. Additionally, students who did not speak the language of the test at home were also excluded[3]. Results of estimation based on this sample are given in the Table 3 below. Comparing to results presented in the Table 2 it is clear that restricting samples to native students diminished the effects of tracking. All estimates of tracking coefficients are closer to zero, however, they remained negative except the case of science.

Table 3. DD estimates based on the sample restricted to native students.

	PIRLS and PISA		**TIMSS 2003 and PISA 2003**			
	2001/2000 Reading	*2001/2003 Reading*	*Mathematics*	*Science*	*Problem solving and mathematics*	*Problem solving and science*
Tracking	-27.62** (12.42)	-16.85* (8.42)	-8.53 (17.18)	0.33 (11.42)	-10.73 (16.93)	-3.17 (16.67)
4TH grade achievement	0.61*** (0.15)	0.44*** (0.12)	0.63*** (0.13)	0.55*** (0.09)	0.69*** (0.13)	0.69*** (0.13)
Constant	203.21** (77.24)	285.07*** (62.22)	184.78** (65.13)	223.29*** (44.16)	155.08** (64.17)	154.48** (64.47)
Adj. R-squared	0.461	0.413	0.613	0.732	0.664	0.662
N	23	20	15	15	15	15

[3] In the case of language problem of multilingual countries arise. Robustness of results were checked by repeating analysis only for students born in the country of the test regardless of the language spoken at home. These results were nearly the same.

4.1.4. Analysis with restrictions on grade and students' age.

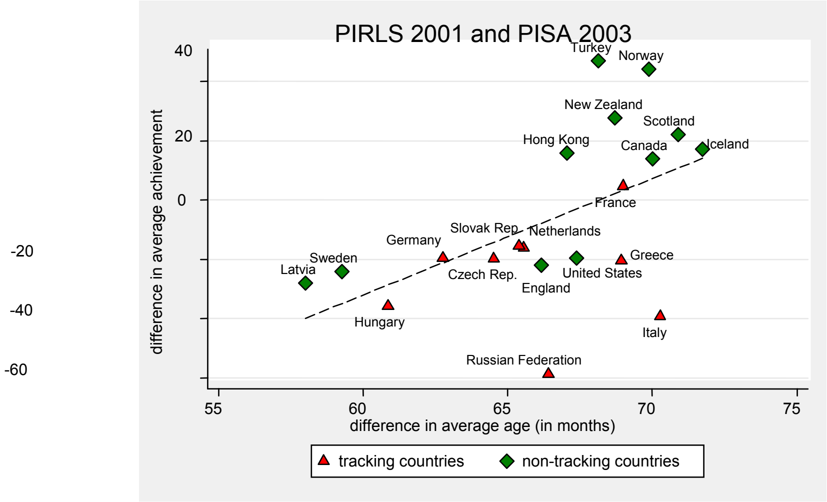
PIRLS and TIMSS differ from PISA in one key assumption. The goal of the latter is to assess achievement of 15-year-olds, so the sample was defined according to students' age, while the former were proposed as a mean to compare achievement of 4th graders regardless of their age. While in PIRLS or TIMSS datasets we have students in the 3rd or 5th grade only in few countries, in PISA tested students were in anything between the 7th and 12th grade. In PIRLS and TIMSS the youngest students tested are 6-year-old and the oldest 15-year-old with important differences in average age between countries. Those differences should be taken into account to make achievement in countries more comparable. This is especially true in the empirical approach used in this study because one could presume that tracking and non-tracking countries systematically differ in the grade-repetition policy which affects the sample of students from secondary schools. Age distributions in countries tested in PIRLS or TIMSS are quite diverse and differ between tracking and non-tracking countries (see Table A2 in the Appendix).

Additionally, many countries sampled students who were still in primary schools. In most cases the number of such students were small, but for example in the Czech Republic almost half of the students tested in PISA 2003 were still in primary education. Having in mind that this country was considered as the tracking one it is obvious that taking into account whole sample assumes that primary school students were already affected by tracking. Even if they were still in the comprehensive school. One could claim that existence of tracking in secondary school can affect students achievement even in primary school but assumptions of this kind should be carefully tested.

To produce comparable estimates of early achievement those differences should be taken into account. Thus, restrictions on grade and age were imposed. We excluded students who were not in the modal grade tested in the particular country or who were older or younger by more than 6 months comparing to the student of average age in each country. In effect students who were still in primary education, but were tested in PISA, were also not considered, because they were not in the modal grades.

Restrictions put on the samples used to estimate country-level average achievement do not guarantee that systematic differences between age in PIRLS/TIMSS and PISA samples of students are not biasing the estimation of tracking effects. Figure 1 shows that the difference between average age of students tested in the 4th grade in PIRLS and students tested in PISA 2003 is strongly related to similar differences in achievement.

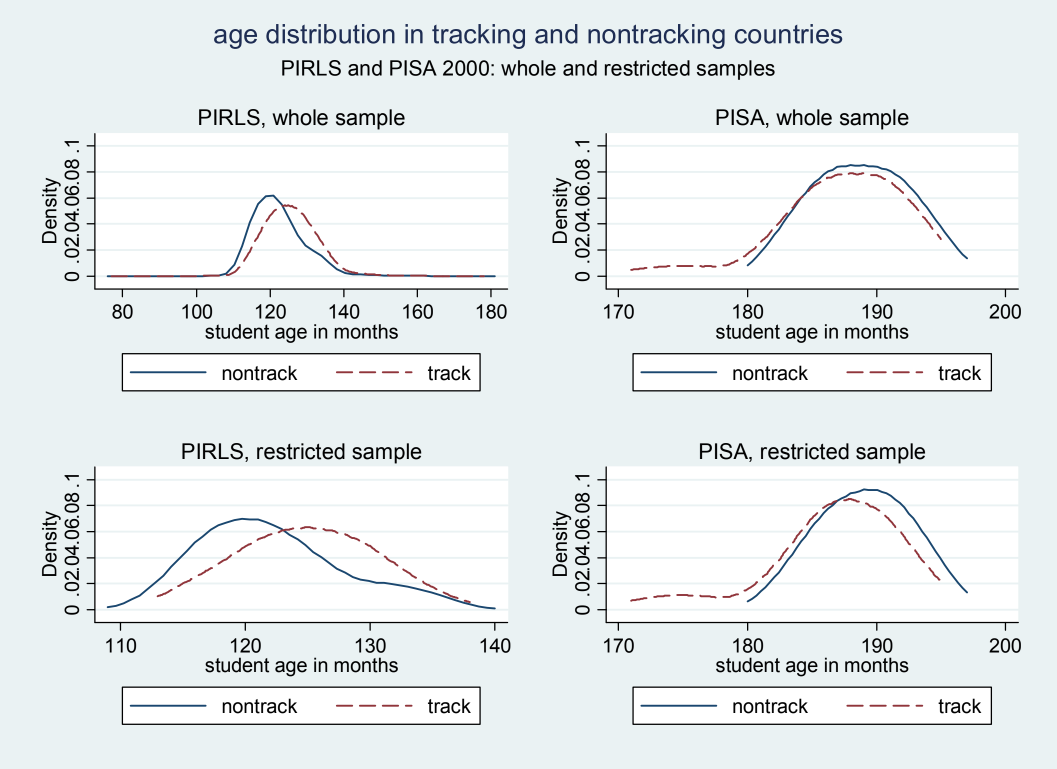
Figure 1. Differences in age and differences in average achievement between PIRLS and PISA 2003.



It is easily visible that the difference in average age of students tested in PIRLS and PISA strongly affects the difference in average achievement between PIRLS and PISA. In some countries average difference in the age of students tested in PIRLS and PISA was bigger by more than one year than in others. To take an extreme example the average age of students tested in PIRLS in Iceland was 116.7 months while in Latvia 132.6 months. Similar numbers for PISA 2003 were 188.4 for Iceland and 190.6 for Latvia. This means that average age difference between PIRLS and PISA 2003 was 71.7 in Iceland while it was only 58 months in Latvia. One could reasonably expect that this will strongly affect differences in achievement because of time-period of learning or maturity effects. It seems that this is the case and that not taking into consideration age differences could bias any estimation based on country-level achievement[4].

There are systematic differences in age distribution between tracking and non-tracking countries. Those are pictured on the graph below for the case of PIRLS and PISA 2000. The upper part of the graph shows age distributions for the whole sample while in the bottom age distribution is sketched after exclusions discussed above were made. Clearly, while restrictions on age and grade make samples from two surveys more comparable there are still systematic differences in average age in PIRLS and PISA between tracking and non-tracking countries. On average students in tracking countries were considerably younger in PIRLS and slightly older in PISA.

Figure 2.



We made an allowance for these differences by running country-level DD regression with two regressors: early achievement and the difference in average age. Results obtained from this regression estimated on the sample restricted to the modal grade and +-6 months around average age in the country are presented in the Table below.

[4] This section points out that not only DD approach is not valid if age differences are not taken into account. The same is true about comparisons of national achievement done in PIRLS or TIMSS. While it was officially claimed that 4th graders achievement is compared regardless of the age it seems that international league Tables produced by PIRLS or TIMSS could be strongly affected by differences in age of tested students. Same is true about any analysis of PIRLS or TIMSS data which do not consider those differences.

Table 4.

	PIRLS/PISA		TIMSS/PISA 2003			
	2000 Reading	2003 Reading	Mathematics	Science	Mathematics/ Problem Solving	Science/ Problem Solving
Tracking	-17.98 (14.21)	-11.49 (10.22)	5.09 (10.56)	13.48 (11.41)	1.90 (8.66)	10.22 (13.54)
4TH grade achievement	0.62*** (0.16)	0.73*** (0.19)	0.61*** (0.09)	0.42*** (0.10)	0.69*** (0.07)	0.63*** (0.11)
Difference in age	2.94** (1.09)	2.77* (1.43)	6.11*** (1.27)	3.18** (1.38)	5.45*** (1.04)	4.77** (1.64)
Constant	15.59 (110.18)	-40.02 (164.31)	-210.49* (105.87)	78.30 (113.96)	-201.44** (86.85)	-127.36 (135.22)
Adj. R-squared	0.505	0.387	0.811	0.604	0.880	0.697
N	23	20	15	15	15	15

Please note positive and significant coefficients of the difference in average age between PIRLS and PISA which shows that discrepancies between countries in the age of testing have to be considered. Comparing these results to those obtained from the full sample with no correction for age and grade we see that in the case of PIRLS and PISA tracking effects are still negative but closer to zero and insignificant. In the case of TIMSS and PISA all coefficients of tracking became positive but still non significant, and coefficients of age difference are highly correlated with outcome proving that correction for age was needed. Results show that correcting for age and grade make earlier conclusions doubtful. It is no longer clear whether there is any visible impact of tracking.

4.1.5. Corrected DD results with age, grade and migrants restrictions

Finally, we estimated DD regression with all restrictions and modifications discussed in points 4.1.3 and 4.1.4 above. Thus, we narrowed the samples to native students, those speaking at home in the language of the test, students from modal grades within the +-6 months brackets around the average age in the tested sample in each country. We also added age difference to regressors. Additionally, we again standardized original scores to have mean 500 and standard deviation 100 in the restricted sample of students in analyzed countries. That was done to assure that excluded students did not affect score distribution in each country and test[5].

The results are quite surprising. Tracking effects in PIRLS and PISA pairs remain negative, but are much closer to zero and insignificant. Estimated coefficients of tracking for TIMSS and PISA pairs are all non-significant and from the practical point of view non distinguishable from zero. In all cases early achievement is reasonably correlated with achievement in PISA, with noticeably smaller coefficient for science. Note also that country-level age differences between PIRLS/TIMSS and PISA entered all regressions significantly with positive coefficients.

Corrected results suggest that there is no clear impact of tracking on mean performance. It seems that earlier results were driven by differences in the sample design between PIRLS, TIMSS, and PISA, mainly by systematic differences in age distributions.

[5] Results were also checked against estimation with scores standardized for the original sample. No visible discrepancies were found.

Table 5. Final DD results obtained for restricted sample and corrected for age differences.

	PIRLS and PISA		**TIMSS 2003 and PISA 2003**			
	2000 Reading	*2003 Reading*	*Mathematics*	*Science*	*Mathematics/ Problem Solving*	*Science/ Problem Solving*
Tracking	-16.36 (13.25)	-15.73 (10.43)	2.18 (14.22)	8.81 (12.77)	-1.93 (12.13)	7.29 (16.62)
4TH grade achievement	0.62*** (0.14)	0.67*** (0.18)	0.59*** (0.11)	0.38*** (0.10)	0.65*** (0.09)	0.57*** (0.13)
Difference in age	3.07*** (1.01)	2.47* (1.40)	5.71*** (1.69)	2.91* (1.54)	4.94*** (1.44)	4.47** (2.00)
Constant	-1.00 (97.97)	0.59 (154.85)	-179.70 (135.84)	106.63 (122.21)	-161.73 (115.90)	-92.10 (159.10)
Adj. R-squared	0.579	0.394	0.696	0.510	0.786	0.584
N	23	20	15	15	15	15

4.2. Effects of tracking on inequalities

In their paper Hanushek and Woessmann suggested that tracking increases inequalities measured by the change in countries dispersion of students scores between primary and secondary schools. They used country-level DD approach again, but mean performance was replaced by standard deviation of scores or other similar statistics (e.g. interquartile range – the difference between the 75th and 25th percentile of each country score distribution). They found that except the PIRLS/PISA 2000 pair all other estimated tracking effects were positive which means that dispersion of scores in secondary schools in tracking countries was higher than in non-tracking countries controlling for the dispersion of scores in primary schools. Regarding the fact that most coefficients were non significant it was argued that great coherence of results proved that tracking increases inequalities. These results are of great interest to for policy-makers and researchers. However, we claim that those results are even more doubtful than results obtained for tracking effects on mean performance. Arguments and additional results are discussed below.

First, there are methodological problems with assessing dispersion scores in PIRLS, TIMSS or PISA. It is well known that with the instruments used to test students knowledge in international tests dispersion of scores is much more difficult to measure than the mean performance (see Koretz et al., 2001). It was also shown that distribution of scores depends on methodological choices made by test organizers (see Brown et al., 2005). While those problems are more important in the case of non OECD countries, especially those low-performing, it is worth checking whether different specifications and statistics used to assess score dispersion produce coherent evidence on tracking effects.

Second, we already discussed the discrepancies between samples of students tested in each country and between sample design of PIRLS, TIMS and PISA. Surely, scores dispersion is heavily affected by the way the sample was constructed, especially if there are differences between age distribution, tested grades, and the proportion of migrants. For example, it seems invalid to assume comparable distributions of scores in the Czech Republic where large part of students tested in PISA were still in primary school, Germany where all students were already in the tracking system, and Norway where all students were still in a comprehensive school. In estimating and comparing scores dispersion it seems crucial to correct for such differences.

Third, Hanushek and Woessmann considered several surveys which test reading, mathematics and science. However, these surveys have different emphasis on subjects and in the case of PISA the main subject changes between years. In PISA 2000 the main domain was reading while in 2003 it was mathematics. It seems valid to compare reading literacy in PIRLS and PISA 2000 but conclusions based on comparisons of reading literacy in PIRLS and PISA 2003 should be treated with great cautious, especially when one concentrates not on the average performance but on much more difficult to estimate statistics like score dispersion. Similarly, comparisons of score dispersion in mathematics in PISA 2003 and TIMSS are more valid then the same comparisons done for science which a minor domain tested less precisely. These facts seem to be overlooked in Hanushek and Woessmann study where examples are focused on reading literacy in PIRLS and PISA 2003. Tracking effects on inequality were found to be positive for these surveys but the negative effects found in a more valid comparison of PIRLS and PISA 2000 were not discussed at length. Other positive effects found in TIMMMS 4th graders and 8th graders comparisons were very close to zero and are not that convincing because of smaller and distinct set of tracking countries[6]. Thus, we claim that two comparisons are the most reliable: PIRLS and PISA 2000, and TIMSS and PISA 2003, because they were focused on the same subjects.

The two Tables below report estimates of tracking effects produced by the same methods as earlier but for standard deviation, interquartile range, and performance at the 10th and 90th percentile. The first Table gives results for whole sample available in PIRLS, TIMSS and PISA, while the second Table employed similar analysis on the restricted sample constructed in the way already described in sections 1.4.3-1.4.5 - limiting the sample to native students, modal grades, students within the +-6 months brackets around the average age in the tested sample in each country. Additionally, we added age difference to the set of explanatory variables in the regressions explaining performance at the 10th and 90th percentile. Again, achievement scores were standardized to have mean 500 and standard deviation 100 after exclusions were made to make country achievement distributions comparable. In this case standardization is especially important because original statistics of dispersion could be heavily affected by excluded students and we claim that they should not be compared among surveys.

Following instructions given in PIRLS, TIMSS and PISA documentation we calculated all statistics separately for each of plausible values and used average of five estimates in the final regression[7]. However, no attempt was made to recalculate the weights for the restricted sample. To make Tables comparable we didn't use original weights even for the whole sample analysis but results obtained with weights were nearly the same. Having that said we understand that proper analysis should be done with weights recalculated for the restricted sample. Nevertheless, it seems unlikely that it could change results in an important way.

No simple conclusion about tracking effects on educational inequalities could be given based on the results presented in the first Table obtained for the whole sample. All estimates are very close to zero, most of them are negative. There is only one positive estimate for PIRLS and PISA 2003 reading literacy, significant at the 10% level. The estimates of tracking for most reliable pairs: reading in PIRLS/PISA 2000 and mathematics in TIMSS/PISA 2003 are not significant and from the practical point of view are non distinguishable from zero. Thus, those results do not support Hanushek and Woessmann claim that DD results suggest strong negative impact of tracking on educational inequalities or dispersion of achievement.

The second Table gives even more striking and confusing results. After restricting the samples to make them more comparable we obtained negative coefficients for all pairs except PIRLS/PISA 2003, but even in this case estimated effects of tracking were very close to zero. Thus, it seems that if we concentrate on native students of the same age and from the modal grade then negative impact of tracking on inequalities disappears. If there is any evidence then in the opposite direction – tracking decreases score dispersion in secondary schools controlling for score dispersion in primary schools.

The choice of statistic measuring score dispersion is not an issue here. Estimates obtained for standard deviations and interquartile range are qualitatively the same. One should note, however, that in most cases correlation between score dispersion in primary and secondary school is very weak which makes DD parametric approach questionable.

Finally, tracking effects were also estimated for the performance at the 10th and 90th percentiles. Based on the full sample estimates one could conclude that there is weak evidence about stronger negative impact on tracking on low-achievers, because all tracking estimates are lower for the 10th than 90th percentile. However, based on the estimates obtained for the restricted sample conclusions could be different because in some

[6] Obviously the number countries who already track 8th grade students is much smaller. These are mainly countries which had German-like structure of education and obviously also share common characteristics other than tracking.

[7] While similar procedures should be undertaken for any statistic based on plausible value they have negligible impact on country averages but play an important role in estimating variance or performance at specific percentiles, especially for countries which are far from the mean of distribution (see OECD, 2005).

cases tracking coefficient is lower for high-achievers. It is hard to imagine why tracking could have stronger negative impact on more able than less able students. Thus, to validate these results additional evidence is needed.

It worth looking at graphs provided below summarizing differences of score distributions between tracking and non-tracking countries based on the restricted sample. It seems that in the case of PIRLS/PISA 2000 and PIRLS/PISA 2003 the middle of scores distribution in tracking countries moved left comparing to non-tracking countries. This shift was much more visible in the case of PIRLS/PISA 2000, however, in both cases we do not observe any differential effects of tracking, namely, we do not see that distribution is "wider" in secondary school than in primary school in tracking countries or that it became more skewed which could be expected having in mind that tracking should mainly affect students who went to non-comprehensive schools. Graphs for TIMSS and PISA data in mathematics and science are even more confusing. There is no clear path of change between tracking and non-tracking countries. It could be claimed that there are important and systematic differences between tracking and non-tracking countries that influence changes in score distribution. In that case simple regression DD method could mistakenly attribute those changes to tracking. Clearly, methods which test for the impact of different policies or time-trends are needed here.

Table 6. Tracking effects on score dispersion. Full sample.

	PIRLS/PISA		**TIMSS/PISA 2003**			
	2000 Reading	*2003 Reading*	*Mathematics*	*Science*	*Mathematics/ Problem Solving*	*Science/ Problem Solving*
Dependent variable: SD of 15-year-olds achievement						
Tracking	-2.39	5.65*	0.58	-0.45	0.51	0.39
	(2.21)	(3.10)	(3.29)	(3.24)	(3.49)	(3.20)
SD of achievement	0.23***	0.33**	-0.21	-0.06	-0.15	-0.14
in the 4TH grade	(0.07)	(0.12)	(0.13)	(0.09)	(0.14)	(0.09)
Constant	72.97***	61.46***	107.69***	98.99***	102.22***	101.30***
	(7.30)	(12.26)	(11.30)	(7.67)	(11.99)	(7.57)
Adj. R-squared	0.343	0.226	0.090	-0.120	-0.029	0.075
Dependent variable: IQR of 15-year-olds achievement						
Tracking	-1.34	9.07*	1.04	-0.85	0.09	0.40
	(3.24)	(4.67)	(4.75)	(4.69)	(4.14)	(3.65)
IQR of achievement	0.25***	0.33**	-0.29**	-0.13	-0.20*	-0.16**
in the 4TH grade	(0.07)	(0.14)	(0.13)	(0.08)	(0.11)	(0.07)
Constant	96.03***	82.62***	155.97***	144.01***	145.51***	140.39***
	(9.62)	(18.33)	(15.08)	(9.73)	(13.14)	(7.58)
Adj. R-squared	0.345	0.190	0.227	0.027	0.112	0.256
Dependent variable: 15-year-olds 10th percentile's achievement						
Tracking	-30.44**	-26.33**	-15.42	-6.22	-15.76	-11.38
	(12.56)	(9.19)	(13.59)	(6.80)	(15.21)	(12.38)
4TH grade 10th percentile's	0.50***	0.34***	0.47***	0.37***	0.49***	0.48***
achievement	(0.11)	(0.11)	(0.09)	(0.04)	(0.10)	(0.08)
Constant	197.23***	258.01***	202.62***	232.67***	190.75***	197.03***
	(41.27)	(39.14)	(34.87)	(16.14)	(39.02)	(29.37)
Adj. R-squared	0.494	0.376	0.643	0.848	0.616	0.734
Dependent variable: 15-year-olds 90th percentile's achievement						
Tracking	-28.75**	-8.49	-3.90	2.90	-6.07	3.72
	(12.14)	(7.50)	(19.11)	(13.75)	(17.76)	(19.83)
4TH grade 90th percentile's	0.65***	0.53***	0.78***	0.75***	0.87***	0.93***
achievement	(0.20)	(0.14)	(0.19)	(0.16)	(0.18)	(0.23)
Constant	226.67*	290.53***	148.49	165.82	89.96	51.61
	(124.11)	(84.32)	(113.36)	(96.04)	(105.35)	(138.52)
Adj. R-squared	0.403	0.458	0.521	0.593	0.620	0.511
N	23	20	15	15	15	15

Table 7. Tracking effects on score dispersion. Restricted sample

	PIRLS/PISA		TIMSS/PISA 2003			
	2000 Reading	*2003 Reading*	*Mathematics*	*Science*	*Mathematics/ Problem Solving*	*Science/ Problem Solving*
Dependent variable: SD of 15-year-olds achievement						
Tracking	-4.76 (3.22)	3.20 (4.24)	-2.95 (4.88)	-4.71 (5.42)	-2.62 (6.08)	-3.49 (5.60)
SD of achievement in the 4TH grade	0.27** (0.10)	0.45** (0.16)	-0.10 (0.18)	-0.08 (0.14)	-0.11 (0.23)	-0.17 (0.14)
Constant	67.61*** (10.16)	52.71*** (16.16)	101.34*** (15.83)	102.64*** (12.43)	101.27*** (19.72)	107.15*** (12.85)
Adj. R-squared	0.317	0.271	-0.120	-0.087	-0.138	-0.034
Dependent variable: IQR of 15-year-olds achievement						
Tracking	-4.30 (5.16)	3.81 (6.02)	-4.26 (6.94)	-8.48 (7.87)	-5.37 (8.49)	-5.92 (7.63)
IQR of achievement in the 4TH grade	0.27** (0.11)	0.46** (0.17)	-0.14 (0.18)	-0.14 (0.13)	-0.17 (0.22)	-0.21 (0.13)
Constant	90.88*** (14.92)	71.33*** (23.26)	142.67*** (20.81)	147.84*** (16.21)	144.91*** (25.43)	149.34*** (15.72)
Adj. R-squared	0.203	0.229	-0.099	-0.015	-0.098	0.046
Dependent variable: 15-year-olds 10th percentile's achievement						
Tracking	-10.14 (16.95)	-16.87 (13.74)	9.51 (11.60)	18.27 (12.48)	7.34 (10.44)	13.76 (13.52)
4TH grade 10th percentile's achievement	0.51*** (0.13)	0.67*** (0.18)	0.39*** (0.07)	0.22** (0.07)	0.44*** (0.07)	0.36*** (0.08)
Difference in age	3.09** (1.27)	3.95** (1.85)	5.26*** (1.35)	2.63 (1.47)	5.12*** (1.22)	4.61** (1.59)
Constant	-10.83 (98.00)	-141.80 (168.21)	-131.08 (100.45)	105.32 (106.65)	-140.60 (90.44)	-75.32 (115.50)
Adj. R-squared	0.464	0.366	0.736	0.466	0.800	0.653
Dependent variable: 15-year-olds 90th percentile's achievement						
Tracking	-15.28 (12.34)	-3.69 (9.12)	8.58 (12.50)	13.32 (12.15)	5.17 (11.88)	15.79 (18.40)
4TH grade 90th percentile's achievement	0.75*** (0.19)	0.54*** (0.16)	0.79*** (0.12)	0.65*** (0.14)	0.91*** (0.12)	0.90*** (0.21)
Difference in age	2.75*** (0.94)	1.00 (1.16)	6.72*** (1.53)	3.90** (1.51)	5.72*** (1.45)	5.36** (2.28)
Constant	-22.73 (140.09)	218.81 (146.64)	-316.12** (142.88)	-38.80 (149.60)	-320.56** (135.78)	-292.04 (226.51)
Adj. R-squared	0.548	0.320	0.773	0.607	0.818	0.553
N	23	20	15	15	15	15

Figure 3. Reading literacy score distribution in tracking and non-tracking countries. PIRLS and PISA 2000. PIRLS and PISA 2003.

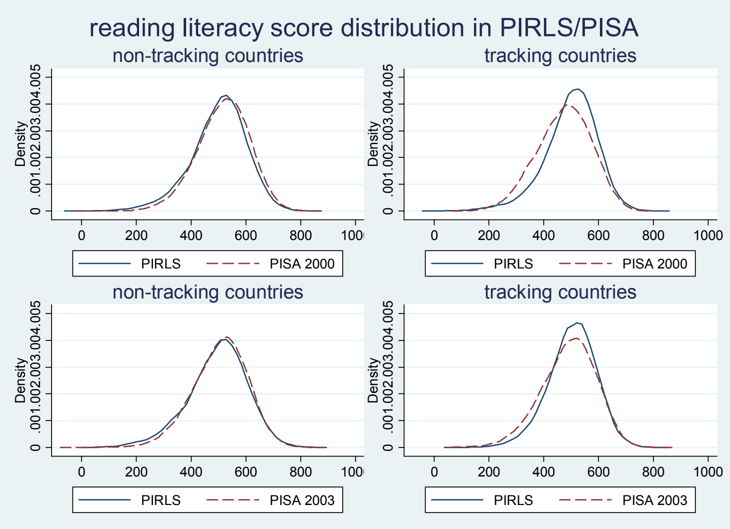
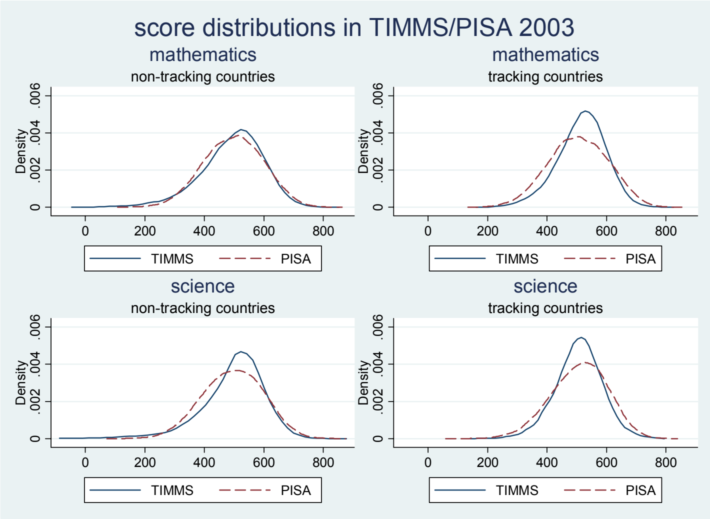


Figure 4. Mathematics and science score distribution in tracking and non-tracking countries. TIIMMS and PISA 2003.



4.3. Do tracking similarly affects students with different family background?

Estimates of average effect of tracking are not very informative. More interesting question is whether different groups of students are similarly affected. We can imagine a situation where on average tracking do not change anything however have strong impact on specific groups, e.g. lowers achievement of disadvantaged groups and increases performance of top-achievers. In this section we checked whether tracking differently affect groups of students who differ in family background which was defined according to parental education and the number of books at home. Parental education was used in the case of PIRLS and PISA after recoding relevant variables into common categories. In the case of TIMSS and PISA the number of books at home was used as a proxy for family background because information about 4th graders parents' education was not collected in TIMSS.

The presumption here was that tracking affects more heavily students with poorer background. Those students more often end in vocational (or other non-comprehensive) schools where educational expectations are lower. They are additionally affected by peers who on average will be less able and less motivated than peers in comprehensive schools where most of the students with privileged family background will end. Obviously, we observed also high-achievers coming from families with poor educational background, however, in most countries such students will end in the vocational track with higher probability than students from well educated families with the same ability or knowledge.

From the other hand, it is hard to imagine that students from well educated families in the tracking country could achieve less in a comprehensive secondary school than similar students in the non-tracking country. If there is any prediction about the impact on these students then it is that tracking will have positive effect on them because of smaller disparities between peers in comprehensive schools in the tracking country. While it is possible in the non-tracking country to have low-achievers in one class in secondary school together with high-achievers that is probably less common in the tracking country where low-achievers were already put to non-comprehensive, often vocational, tracks. Thus, we expected that tracking will have positive or negligible impact on high-achievers. This way we could also check validity of our results. The negative impact of tracking on all students regardless of their family background could be a sign of invalid statistical model.

Results of estimation are presented in the two Tables below, separately for reading in PIRLS/PISA and mathematics, sciences, and problem solving in TIMSS/PISA. There are only results obtained for the restricted sample of native students in modal grades and age within the +-6 months brackets around the average age in the country's sample (restrictions similar to those made in section 1.4.3-1.4.5). Results for the whole sample and more detailed categories are presented in the Appendix (Tables A7-A8).

Results for the students with missing data on parents' education or the number of books at home are also given. These data are not missing at random and in some countries even half of the sampled students didn't provide that information. In the case of PIRLS United States were excluded from the analysis because there was no data on parental education provided in the dataset. Thus, one should keep in mind when looking at the Table that similar students could be differently classified in PIRLS or TIMSS and PISA with additional bias owing to huge differences in response rates between surveys in some countries (e.g. Argentina). It is important also to notice that only PIRLS data were from questionnaires addressed to parents while data in PISA and TIMSS were reported by students themselves. This means that on average we have more missing data on education in PIRLS but data in PISA are less reliable and probably biased upward. Thus, in many cases in PISA category "tertiary" contains students whose parents were in fact less educated and should be matched to students from other categories. Thus, if assumption that tracking more heavily affects students from less educated families is correct than estimates of tracking effects for category "tertiary" could be biased downward.

Results presented in the Tables 8 and 9 suggest that effects of tracking depend on students family background. These results should be treated as preliminary and analyzed with cautions because none of estimates of tracking was significant. However, results are coherent and in line with hypotheses stated above. In PIRLS/PISA pairs tracking effects for students whose parents have higher education are much closer to zero than negative estimates for students with less educated parents. In TIMSS/PISA pairs estimates for the group of students with the lowest number of books at home are negative or very close to zero, while estimates for other groups are positive. Science is an exception here with positive estimates for all groups.

One could elaborate slightly more on these data to formally test whether differences between groups are significant, however, similar hypotheses will be stated and tested in the next section with micro-data. To sum up, this preliminary analysis suggest that tracking affects more heavily students with poorer background and that we could expect even positive effects for students from well educated families.

Table 8. Parental education and tracking effects (sample of native students corrected for age and grade).

Parental education	PIRLS/2001 and PISA 2000				PIRLS/2001 and PISA 2003			
	Tertiary	*Upper/ post secondary*	*Lower-secondary/ Primary*	*Missing*	*Tertiary*	*Upper/ post secondary*	*Lower-secondary/ Primary*	*Missing*
Tracking	-8.93 (13.57)	-21.10 (15.82)	-23.40 (18.01)	-21.53 (20.29)	-9.89 (12.07)	-20.59* (11.45)	-31.98* (15.47)	-35.26** (15.27)
4^{TH} grade achievement	0.67*** (0.17)	0.64*** (0.20)	0.70*** (0.18)	0.60*** (0.21)	0.44* (0.21)	0.63*** (0.18)	0.65*** (0.20)	0.75*** (0.22)
Difference in age	3.36*** (1.02)	3.08** (1.17)	3.87*** (1.34)	3.43** (1.50)	1.56 (1.53)	1.84 (1.42)	3.36 (1.96)	3.99* (2.02)
Constant	-56.91 (118.05)	-15.03 (127.65)	-102.88 (117.75)	-63.81 (139.01)	181.62 (181.90)	64.31 (151.38)	-56.40 (178.73)	-181.24 (201.28)
Adj. R-squared	0.534	0.454	0.554	0.399	0.077	0.366	0.410	0.405
N	22	22	22	22	19	19	19	19

Table 9. The number of books at home and tracking effects (sample of native students corrected for age and grade).

How many books do you have at home?	**Mathematics**				**Science**			
	0-25	*26-100*	*100+*	*missing*	*0-25*	*26-100*	*100+*	*missing*
Tracking	-6.35 (16.78)	-1.87 (15.72)	5.83 (13.43)	-18.24 (14.23)	5.15 (18.17)	9.81 (15.10)	10.38 (12.31)	-11.81 (13.77)
4^{TH} grade achievement	0.59*** (0.13)	0.69*** (0.14)	0.72*** (0.12)	0.55*** (0.11)	0.31* (0.15)	0.46*** (0.15)	0.54*** (0.11)	0.35*** (0.10)
Difference in age	6.67*** (1.96)	7.18*** (1.87)	6.97*** (1.64)	4.08** (1.59)	3.27 (2.16)	4.08** (1.83)	4.35** (1.51)	0.43 (1.57)
Constant	-272.30 (156.49)	-345.12* (157.88)	-319.01** (140.10)	-74.33 (122.00)	82.96 (172.50)	-25.92 (157.23)	-52.81 (128.46)	261.51** (117.82)
Adj. R-squared	0.634	0.681	0.751	0.671	0.179	0.419	0.637	0.393

How many books do you have at home?	**Mathematics/problem solving**				**Science/problem solving**			
	0-25	*26-100*	*100+*	*missing*	*0-25*	*26-100*	*100+*	*missing*
Tracking	-11.80 (14.92)	-7.17 (12.72)	4.07 (11.02)	-35.67* (17.87)	-1.99 (20.20)	5.58 (19.35)	12.87 (16.52)	-26.28 (20.10)
4^{TH} grade achievement	0.64*** (0.12)	0.79*** (0.11)	0.78*** (0.10)	0.62*** (0.13)	0.54*** (0.16)	0.70*** (0.19)	0.70*** (0.15)	0.54*** (0.15)
Difference in age	6.16*** (1.75)	6.46*** (1.51)	6.10*** (1.35)	2.70 (2.00)	5.68** (2.41)	5.93** (2.34)	5.50** (2.03)	2.18 (2.29)
Constant	-264.04* (139.13)	-347.60** (127.78)	-297.54** (115.02)	-5.94 (153.22)	-185.73 (191.80)	-269.86 (201.50)	-216.23 (172.43)	57.81 (171.97)
Adj. R-squared	0.709	0.796	0.833	0.597	0.444	0.508	0.617	0.467

5. Results for non parametric DD approach

The way tracking effects were estimated by Hanushek and Woessmann and in preceding sections was based on simple linear regression where the dependent variable was PISA country-level statistic and among independent variables was primary school country-level statistic. However, in many regressions presented above the R-square was quite low and statistics of early achievement were weakly correlated with similar measures in the secondary school. That was especially true when several statistics of dispersion were employed (standard deviation or interquartile range). In this case, one could question this regression approach which assumes that there is a strong and linear relation between primary and secondary school statistics. In this section we employed fully non parametric approach which simply compares difference in achievement growth between tracking and non-tracking countries. This was done in two ways. First, with country-level data. Second, with individual student-level data.

Only reading literacy in PIRLS and PISA 2000 and mathematics in TIMSS 2003 and PISA 2003 were considered. These are two pairs of surveys focused on the detailed measurement of achievement in the same subject. Thus, they are the most reliable source of information about individual achievement in reading and mathematics. All regressions were estimated on the full sample with weights produced by survey organizers. However, in student-level regressions we adjusted individual weights to have the same sum for every country making results comparable to those with country-level data. Results obtained with unadjusted weights or even without any weights were substantially the same.

Consider country-level data first. We estimated regression equation [2] (see section 2) where in this case the dependent variable was a country-level mean performance on one of the surveys and among dependent variables there were a dummy indicating whether results are from primary or secondary school (1 for secondary), a dummy indicating tracking countries and interaction term which equals 1 for secondary school test in the tracking country. Thus, β in equation [2] is a key parameter of interest. The results are presented in the Table 10 below.

Table 10. Results for non-parametric DD approach with country-level data.

	PIRLS and PISA 2000 *Reading literacy*	**TIMSS and PISA 2003** *Mathematics*
Time (PISA=1)	13.73 (16.53)	2.08 (22.40)
Tracking countries dummy	5.72 (16.90)	21.60 (27.43)
Time * Tracking	-36.37 (23.90)	-16.90 (38.79)
Constant	493.88*** (11.69)	491.70*** (15.84)
Adj. R-square	0.011	-0.087
N	46	30

Additionally, individual data were explored to estimate similar equation with some extensions possible with student level observations. The goal here was to check whether tracking effects are heterogeneous by separating tracking impact for group of students with less advantageous family background. Moreover, using student-level data opened the possibility of controlling for such variables like gender, migrant status, grade and individual's age.

In practice we estimated several regressions with different sets of covariates and tracking effects tested. Results for all of them are presented in the Tables 11 and 12, for reading in PIRLS and PISA 2000, and for mathematics in TIMSS and PISA 2003, respectively. In all regressions students' gender and dummy indicating whether she or he was born in the country of test or speaks the language of test at home were included together with interaction term with time. This way we partially controlled for different definitions, response rates or even measurement related bias between PIRLS or TIMSS and PISA. For example, girls had on average about 20 points more in reading in PIRLS, while 32 points more in reading in PISA. Regardless slightly different samples and scores standardization estimated coefficient for gender was 20.72 and for interaction with time was 11.24 showing almost the same gender gap as in official PIRLS and PISA reports (see Mullis et al., 2003, page 39; OECD, 2001, page 124). However, we did not add interaction terms of individual characteristics with treatment dummy (i.e., with time and tracking dummies) assuming that those characteristics were similarly defined and measured in all countries participating in a specific survey. Nevertheless, we checked results obtained with treatment interaction terms to find them almost the same.

Results obtained for the regression with these individual characteristics, time, tracking and treatment dummies are in the column (1) of Tables 11 and 12. Negative effects of tracking (interaction term Time*Tracking) are similar to those estimated with country-level data (see Table 10). However, standard errors are much smaller and tracking estimates are highly significant. Those standard errors were computed analytically correcting for clustering at the school level. We tried also bootstrap and jackknife estimates of standard errors but they were only slightly different and did not change any of the conclusions. Again, all regressions were run only with first plausible value provided by survey organizers and were not corrected for randomness in plausible values imputation. However, such correction was too small to modify our conclusions.

In columns (2) - (4) results obtained from regression where tracking effects where estimated separately for group of students with lower family background are presented. Those are called second-order treatment effects (see Meyer, 1995) or difference-in-differences-in-differences estimates (see Gruber, 1994). Basically, we tested our hypothesis that lower background students could be affected by tracking while similar effects for other students should be negligible assuming that there is no other systematic influence on achievement scores in tracking comparing to non-tracking countries. Lower background was arbitrarily defined as not having a parent with tertiary education (PIRLS/PISA) or having at most 100 books at home (TIMSS and PISA). Results with slightly different thresholds were similar. It should be emphasized that we have a "fuzzy" treatment here partly because some students with advantageous family background can end up in vocational school and partly because of measurement error in self-declared variables defining affected groups. Nevertheless, if there is any differential impact of tracking according to family background of students we should be able to detect it.

To control for systematic differences between tracking and non-tracking countries as well as between students defined as "low background" and others we controlled changes in achievement of lower background students among surveys and between two groups of countries considered. Thus, we interacted all levels of parental education or number of books at home with time dummy and estimated interaction term of lower background and tracking dummy. This way we controlled not only for any systematic differences in samples tested in two surveys, but also for differences in measurement or definitions, or general correlation between socio-economic status of families and students achievement growth.

The coefficient of interest is now for the "Low background * Tracking * Time" interaction term. Estimates of this coefficient presented in column (2) suggest that differential impact of tracking on the lower background group of students is negligible, both in PIRLS/PISA and TIMSS/PISA comparisons. To control for age and grade differences between surveys and countries we added age and grade dummies. Those results are presented in column (3). Still tracking impact on lower background students is not significant. Finally, we replaced tracking dummy with full set of country dummies estimating country fixed effects. Those results are presented in column (4). In this case both coefficients are negative, but in the case of PIRLS/PISA not significant. Nevertheless, estimated differential impact on lower background group was in all cases very close to zero suggesting that if there is any impact on that group is almost negligible from the practical point of view.

It is worth to have look at distribution of scores in tracking and non-tracking countries within groups of students with similar family background. This is given in Figure 5 for the PIRLS and PISA 2000 reading literacy scores. Note that regardless of the parental education distribution of students scores in tracking countries is shifted to the left in PISA comparing to tracking countries. Only in the missing data category students from both groups of countries scored lower (in relation to other students). These graphs evidently show that there are systematic differences between tracking and non-tracking countries which almost similarly affect negatively all students in tracking countries. It is obviously possible that those differences are not related to tracking policy only. Similar effects estimated for students with advantageous family background, who are probably in comprehensive schools alike their colleagues in non-tracking countries, suggest that earlier estimates of tracking effects were confounded by other factors. While an interesting research question is what factors have such a negative impact on the group of tracking countries it seems dubious to attribute those effects to tracking alone.

Table 11. Results for non-parametric DD approach with student-level data. Reading literacy in PIRLS 2001 and PISA 2000.

	PIRLS 2001 PISA 2000 *Reading literacy*			
	(1)	(2)	(3)	(4)
Time (PISA=1)	-18.88*** (2.52)	-17.89*** (2.70)	-145.08*** (51.46)	-125.14** (51.28)
Tracking countries dummy	4.88** (2.00)	9.88*** (2.53)	15.50*** (2.49)	
Time * Tracking	**-34.62*** (3.25)**	**-36.99*** (3.80)**	**-20.69*** (3.74)**	**-20.42*** (3.29)**
Gender (1 = girls)	20.72*** (0.90)	20.72*** (0.90)	20.58*** (0.89)	20.48*** (0.86)
Gender * Time	11.24*** (1.54)	11.25*** (1.55)	8.11*** (1.47)	9.03*** (1.30)
Born outside the country	-30.31*** (1.87)	-30.37*** (1.87)	-31.10*** (1.79)	-31.50*** (1.60)
Born outside the country * Time	20.49*** (3.44)	20.52*** (3.44)	35.32*** (3.31)	28.00*** (2.66)
Different language at home	-22.90*** (3.31)	-22.84*** (3.31)	-20.20*** (3.28)	-22.52*** (3.16)
Different language at home * Time	-15.78*** (5.34)	-15.65*** (5.33)	-18.40*** (5.13)	-15.28*** (3.95)
Parents have maximum post/upper secondary education	-41.24*** (1.26)	-38.08*** (1.93)	-37.23*** (1.86)	-41.03*** (1.68)
Parents have maximum lower-secondary education	-73.19*** (2.17)	-69.96*** (2.65)	-71.23*** (2.58)	-77.55*** (2.22)
Parents have primary or no education	-124.76*** (4.36)	-121.88*** (4.71)	-119.58*** (4.58)	-102.59*** (3.39)
Time * Parents have maximum post/upper secondary education	12.92*** (1.73)	12.01*** (2.53)	13.73*** (2.41)	18.63*** (2.17)
Time * Parents have maximum Lower-secondary education	25.18*** (3.25)	24.05*** (3.84)	31.65*** (3.62)	31.23*** (3.00)
Time * Parents have Primary or no education	41.44*** (5.56)	40.43*** (6.02)	53.62*** (5.63)	41.53*** (4.08)
Low background * Tracking		-6.77** (2.64)	-7.71*** (2.58)	-1.58 (2.31)
Low background * Tracking * Time		**2.29 (3.64)**	**3.18 (3.43)**	**-4.19 (3.08)**
Age and grade dummies			YES	YES
Country Fixed Effects				YES
Constant	538.57*** (1.82)	536.42*** (2.06)	494.95*** (16.16)	441.53*** (16.47)
Adj. R-square	0.125	0.125	0.177	0.272
N	186396	186396	185324	185324

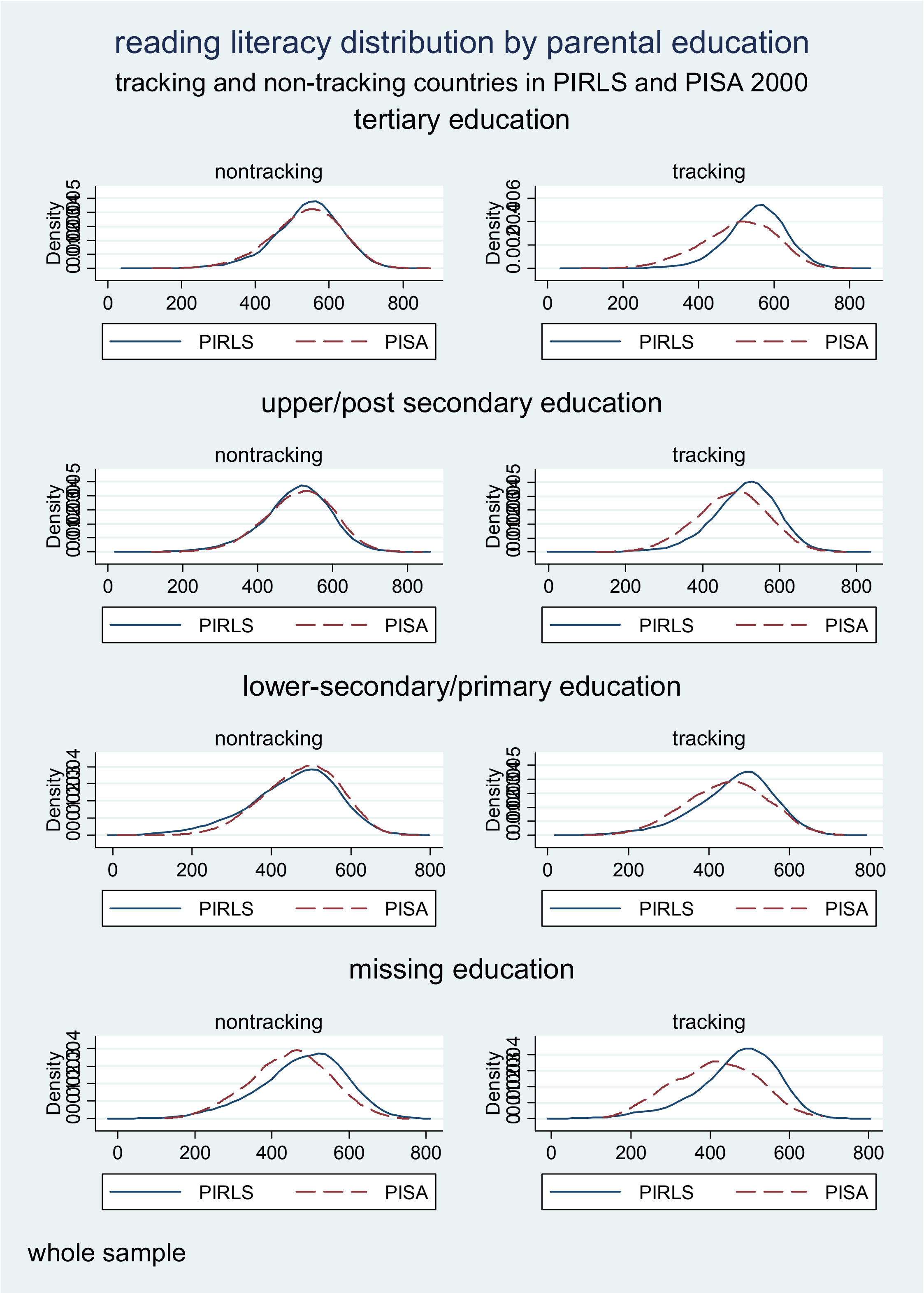
* p<0.10, ** p<0.05, *** p<0.01

Table 12. Results for non-parametric DD approach with student-level data. Mathematics in TIMSS 2003 and PISA 2003.

	TIMSS 2003 PISA 2003 *Mathematics*			
	(1)	(2)	(3)	(4)
Time (PISA=1)	-0.06 (5.17)	-0.60 (5.31)	0.00 (0.00)	0.00 (0.00)
Tracking countries dummy	27.44*** (2.39)	23.36*** (2.59)	20.40*** (2.46)	
Time * Tracking	**-24.64*** (3.72)**	**-25.07*** (3.91)**	**-13.77*** (3.89)**	**-6.83** (3.40)**
Gender (1 = girls)	-6.82*** (0.92)	-6.84*** (0.92)	-6.94*** (0.90)	-6.66*** (0.80)
Gender * Time	-4.65*** (1.56)	-4.62*** (1.56)	-7.47*** (1.54)	-7.26*** (1.32)
Born outside the country	-15.85*** (2.43)	-15.84*** (2.43)	-13.03*** (2.48)	-28.41*** (1.69)
Born outside the country * Time	30.50*** (3.23)	30.62*** (3.23)	40.95*** (3.48)	35.69*** (2.52)
Different language at home	-54.74*** (4.86)	-54.66*** (4.85)	-50.88*** (4.62)	-31.99*** (3.83)
Different language at home * Time	35.96*** (5.79)	35.73*** (5.79)	25.34*** (5.58)	11.28** (4.77)
11-25 books	38.21*** (3.22)	38.05*** (3.20)	35.74*** (3.03)	26.36*** (1.77)
26-100 books	65.63*** (3.59)	65.53*** (3.57)	62.31*** (3.33)	49.29*** (1.87)
101-200 books	76.00*** (3.84)	77.79*** (4.22)	74.90*** (3.95)	65.76*** (2.19)
More than 200 books	76.36*** (3.94)	78.14*** (4.32)	76.15*** (4.04)	67.96*** (2.30)
Time * 11-25 books	-22.05*** (3.72)	-22.09*** (3.69)	-26.28*** (3.42)	-11.92*** (2.26)
Time* 26-100 books	-14.35*** (4.22)	-14.51*** (4.19)	-27.09*** (3.71)	-13.95*** (2.42)
Time* 101-200 books	-3.50 (4.59)	-2.68 (5.11)	-19.06*** (4.49)	-9.16*** (2.93)
Time * More than 200 books	23.53*** (4.76)	24.34*** (5.25)	5.65 (4.60)	15.27*** (3.06)
Low background * Tracking		5.84** (2.64)	7.88*** (2.56)	11.56*** (2.26)
Low background * Tracking * Time		**2.49 (3.78)**	**-4.92 (3.58)**	**-13.29*** (3.18)**
Age and grade dummies			YES	YES
Country Fixed Effects				YES
Constant	446.19*** (4.24)	445.68*** (4.33)	480.92*** (5.66)	499.17*** (5.96)
Adj. R-square	0.110	0.110	0.175	0.336
N	148712	148712	147960	147960

* p<0.10, ** p<0.05, *** p<0.01

Figure 5. Distribution of reading literacy achievement scores in tracking and non-tracking countries by parental education. PIRLS and PISA 2000.



6. Summary and conclusions

The goal of this paper was to discuss the difference-in-differences approach to estimate tracking effects based on the PIRLS, TIMSS and PISA data which allows international comparisons of student achievement in reading, mathematics, science and problem solving. We extended the seminal work of Hanushek and Woessmann who claimed that tracking has negative impact on educational inequalities and at least no positive impact on mean performance. Using individual data we corrected the samples of students tested in TIMSS, PIRLS, and PISA to make them more comparable. It was shown that there are crucial differences between surveys that could biased DD results. In fact, employing Hanushek and Woessmann method we estimated tracking effects on the samples corrected for age, grade, migrant status, and difference between the language of the test and spoken at home. While results for the non-restricted sample were qualitatively the same as those obtained earlier we found that estimates based on the restricted sample were quite different. No evidence that tracking increases educational inequalities was found. Additionally, we found some evidence on the negative impact of tracking on reading literacy but estimates for mathematics, science and problem solving were very close to zero.

Finally, we tried different DD approach relaxing the assumption of linear relation between countries achievement in primary and secondary education. Using individual data we were also allowed to control for student characteristics and estimate tracking effects more precisely. The most simple DD estimator suggested that tracking lowers mean performance. However, we found no evidence that tracking affects more heavily students with less advantageous family background which was expected to be the case. It seems that the difference between primary and secondary school achievement in tracking countries is lower but this was observed for all groups of students. Arguments provided in the paper suggested that tracking was confounded with other factors which in systematic way affected all students in tracking countries in this case. Thus, earlier work based on the same data which suggested that tracking increases inequalities and has no positive impact on mean performance should be revised. No such evidence was found here and it is doubtful whether simple difference-in-differences approach is able to estimate unbiased tracking effects without further adjustments and methodological developments.

References

Brown, Giorgina, Micklewright, John, Schnepf, Sylke V. and Waldmann, Robert (2005) Cross-National Surveys of Learning Achievement: How Robust are the Findings? Southampton, UK, Southampton Statistical Sciences Research Institute, 36pp. (S3RI Applications and Policy Working Papers, A05/05)

Bruce D. Meyer, 1995. "Natural and Quasi-Experiments in Economics," Journal of Business & Economic Statistics, vol 13, April 1995, pp 151-162.

Eric A. Hanushek & Ludger Wössmann, 2006. "Does Educational Tracking Affect Performance and Inequality? Differences- in-Differences Evidence Across Countries," Economic Journal, Royal Economic Society, vol. 116(510), pages C63-C76, 03.

Gruber, Jonathan, 1994. "The Incidence of Mandated Maternity Benefits," American Economic Review, American Economic Association, vol. 84(3), pages 622-41, June.

Koretz, D., McCaffrey, D., and Sullivan, T. (2001, September 14). Predicting variations in mathematics performance in four countries using TIMSS. Education Policy Analysis Archives, 9(34).

Martin, M.O., Mullis, I.V.S., & Chrostowski, S.J. (Eds.)(2004), TIMSS 2003 Technical Report, Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M.O., Mullis, I.V.S., & Kennedy, A.M. (Eds.) (2003), PIRLS 2001 Technical Report, Chestnut Hill, MA: Boston College.

Mullis Ina V.S., Martin Michael O., Gonzalez Eugene J., Kennedy Ann M, 2001. “PIRLS 2001 International Report”, International Study Center, Lynch School of Education, Boston College.

OECD (2002), *PISA 2000 Technical Report,* OECD, Paris

OECD (2005), *PISA 2003 Technical Report,* OECD, Paris

OECD, 2001. Knowledge and Skills for Life: First Results from PISA 2000.

Appendix

Table A1. Standardized mean achievement on PIRLS, TIMSS and PISA in the analyzed sample.

	PIRLS 2001	PISA 2000	PIRLS 2001	PISA 2003	TIMSS 2003	PISA 2003	TIMSS 2003	PISA 2003	PISA 2003
country	*Reading literacy*		*Reading literacy*		*Mathematics*		*Science*		*Problem Solving*
Argentina	362.7	427.3							
Australia					487.2	521.2	511.7	517.0	526.0
Belgium (Flemish Community)					544.1	550.3	509.2	520.8	543.4
Bulgaria	529.3	439.3							
Canada	521.3	542.0	518.8	532.6					
Czech Republic	512.0	499.8	508.9	489.2					
England	532.4	531.2	530.7	508.7	522.8	503.8	532.7	510.7	504.8
France	497.1	512.8	492.9	497.6					
Germany	514.8	492.3	511.9	492.3					
Greece	495.8	482.2	491.6	471.2					
Hong Kong	500.5	533.2	496.6	512.3	570.5	547.3	535.1	531.5	544.2
Hungary	520.1	488.3	517.5	481.8	519.9	487.0	521.4	495.1	497.1
Iceland	480.9	514.9	475.6	492.7					
Israel	476.5	460.8							
Italy	516.9	495.7	514.1	474.9	491.7	462.6	506.3	478.3	465.4
Japan					559.3	531.1	536.1	539.6	543.5
Latvia	521.8	466.7	519.4	491.4	527.9	480.3	523.3	480.9	478.5
Macedonia	390.7	382.1							
Netherlands	534.1	539.6	532.5	516.3	532.9	534.8	516.4	516.3	516.3
New Zealand	501.8	536.5	497.9	525.6	481.6	520.4	510.6	512.8	529.0
Norway	464.0	513.3	457.5	501.5	435.5	492.1	453.4	476.0	485.8
Romania	480.0	436.9							
Russian Federation	500.6	470.3	496.7	437.9	523.4	465.4	517.6	481.1	474.5
Scotland	500.9	533.4	497.0	519.0	478.1	520.8	491.6	505.7	521.7
Slovak Republic			483.3	467.7					
Sweden	542.7	524.2	541.8	517.6					
Tunisia					312.9	355.7	289.8	376.2	340.2
Turkey			389.6	436.6					
United States	518.7	512.4	516.1	496.5	508.7	479.8	527.7	483.1	473.3

Note : Own calculations based on the average of student level plausible values in each subject and weights provided by survey organizers. All scores were standardized to have mean 500 and standard deviation 100 in the sample of considered countries.

Figure A1. Reading literacy score distribution in tracking countries: PIRLS and PISA 2000.

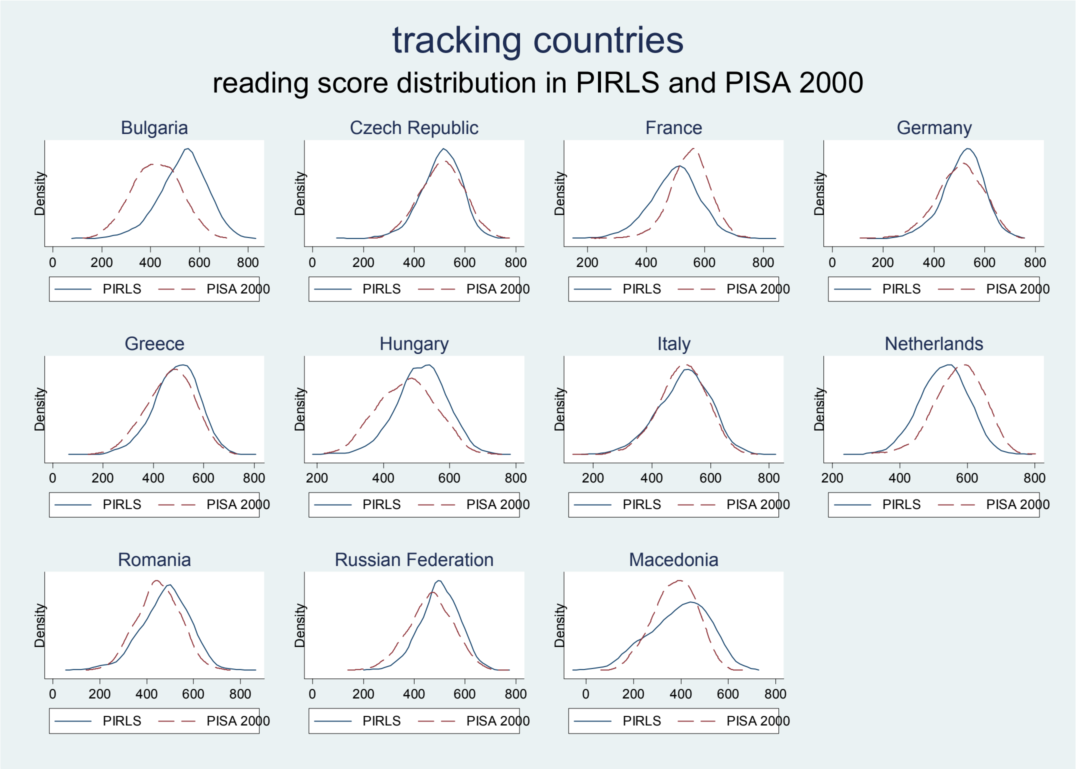


Figure A2. Reading literacy score distribution in non-tracking countries: PIRLS and PISA 2000.

XXXXXXXXXXXXXXXXXXXXXX

Figure A3. Mathematics literacy distribution in tracking countries – TIIMMS and PISA 2003.

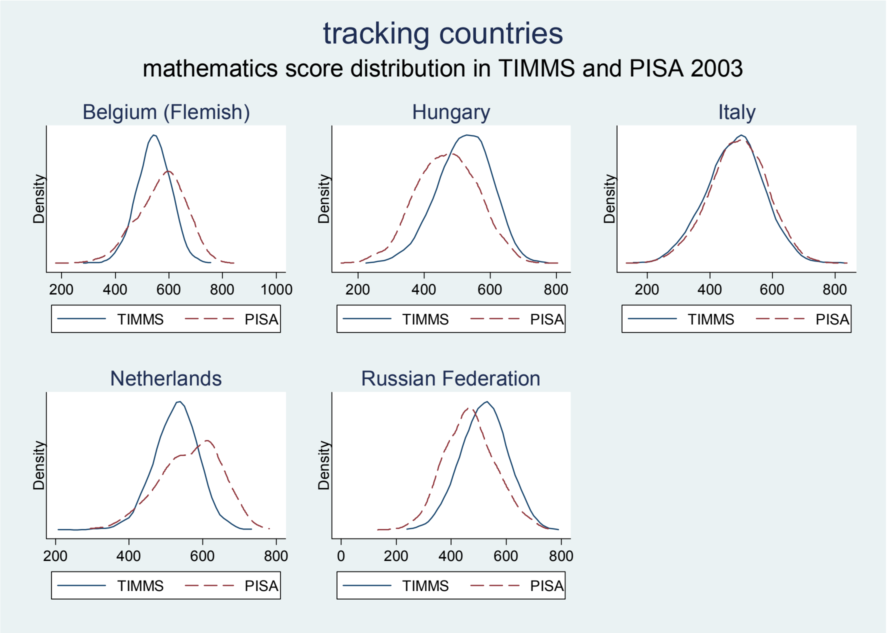


Figure A4. Mathematics literacy distribution in non-tracking countries – TIIMMS and PISA 2003.

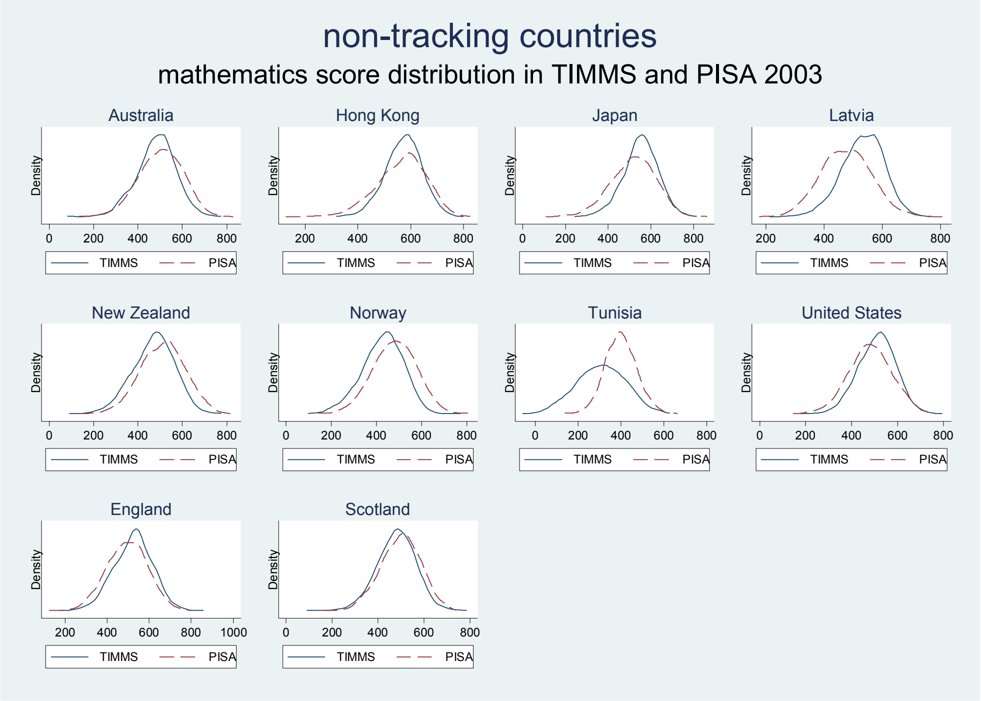


Table A2. Basic statistics and description of student characteristics used in the analysis (not weighted).

NT – non-tracking countries *T – tracking countries*		PIRLS and PISA 2000				TIMSS and PISA 2003			
		PIRLS		PISA 2000		TIMSS		PISA 2003	
		NT	*T*	*NT*	*T*	*NT*	*T*	*NT*	*T*
Gender	% girls	50	49	51	51	50	49	50	50
Born outside the country	%	19	9	9	4	13	8	8	5
Different language at home	%	6	3	8	7	3	2	5	3
Highest level of parents' education	% tertiary	23	18	49	34	No data in TIMSS			
	% post/upper secondary	35	46	32	49				
	% lower-secondary	8	13	10	10				
	% primary or none	5	3	5	4				
	% missing	29	19	5	3				
Age	in months	122.4	125.5	188.8	187.2	122.3	122.6	189.5	189.2
Grade		4.12	3.94	9.9	9.5	4.0	4.0	9.9	9.6
How many books do you have at home?	% 0 – 10 books					13	10	11	7
	% 11-25 books					20	25	14	14
	% 26-100 books					32	35	28	30
	% 101-200 books					16	15	18	20
	% more than 200 books					14	13	26	27
	Missing					4	2	2	2

Table A3. PIRLS and PISA 2000. Country-level DD based on reading literacy subscales.

	PIRLS and PISA 2000 reading literacy subscales PIRLS: 1) reading to acquire and use information; 2) reading for literary experience PISA 2000: 3) retrieving information; 4) interpreting texts; 5) reflection and evaluation						**PIRLS subscales and PISA 2003 overall reading literacy**	
	Information[1] Retrieving[3]	*Experience[2] Retrieving*	*Information Interpreting[4]*	*Experience Interpreting*	*Information Reflection[5]*	*Experience Reflection*	*Information Reading*	*Experience Reading*
track	-34.33** (12.99)	-30.56** (12.84)	-34.19*** (11.88)	-30.49** (11.60)	-43.44*** (11.81)	-39.93*** (11.49)	-28.50*** (8.00)	-24.88*** (7.97)
PIRLS	0.66*** (0.16)	0.65*** (0.15)	0.66*** (0.14)	0.65*** (0.14)	0.62*** (0.14)	0.62*** (0.14)	0.49*** (0.12)	0.47*** (0.12)
Cons	180.35** (77.78)	183.75** (75.88)	184.84** (71.14)	184.95** (68.56)	204.19*** (70.68)	202.96*** (67.93)	261.18*** (61.74)	273.39*** (60.53)
R^2	0.486	0.493	0.528	0.545	0.555	0.574	0.528	0.514
N	23	23	23	23	23	23	20	20

Table A4. TIMSS 2003 and PISA 2003: country-level DD with all mathematics subscales.

TIMSS 2003 mathematics content scales and PISA 2003 mathematics subscales

TIMSS content scales: (alg) – algebra; (dap) - data reproduction, analysis, probability; (fns) - fractions + number sense; (geo) – geometry.; (mea) measurement. **Cognitive scales:** (app) – applying; (kno) – knowing; (rea) – reasoning.

PISA: m1 - space and shape; m2 - change and relationship; m3 – uncertainty; m4 - quantity

	m1	*m1*	*m1*	*m1*	*m1*	*m2*	*m2*	*m2*	*m2*	*m2*
	alg	*dap*	*fns*	*geo*	*mea*	*alg*	*dap*	*fns*	*geo*	*mea*
track	-18.03	-2.44	-22.02	-8.90	-18.01	-13.85	2.90	-16.96	-3.83	-13.91
	(16.12)	(12.39)	(17.40)	(13.73)	(13.41)	(18.29)	(13.11)	(20.95)	(16.09)	(15.33)
TIMSS	0.61***	0.57***	0.67***	0.66***	0.60***	0.67***	0.64***	0.70***	0.71***	0.66***
	(0.12)	(0.08)	(0.15)	(0.11)	(0.09)	(0.14)	(0.09)	(0.18)	(0.13)	(0.11)
Cons	196.14***	212.98***	167.76**	168.44***	199.87***	169.46**	180.03***	154.07	142.04**	172.46***
	(60.72)	(42.27)	(72.55)	(54.97)	(46.90)	(68.90)	(44.72)	(87.33)	(64.44)	(53.63)
R^2	0.620	0.758	0.573	0.709	0.735	0.600	0.779	0.494	0.674	0.716
N	15	15	15	15	15	15	15	15	15	15

	m3	*m3*	*m3*	*m3*	*m3*	*m4*	*m4*	*m4*	*m4*	*m4*
	alg	*dap*	*fns*	*geo*	*mea*	*alg*	*dap*	*fns*	*geo*	*mea*
track	-28.15	-13.61	-30.80	-19.64	-28.55	-9.23	5.17	-12.16	-0.81	-9.34
	(22.56)	(17.32)	(24.57)	(20.08)	(20.23)	(16.08)	(12.28)	(18.04)	(13.82)	(13.51)
TIMSS	0.59***	0.60***	0.62**	0.65***	0.59***	0.57***	0.54***	0.60***	0.62***	0.56***
	(0.17)	(0.12)	(0.21)	(0.16)	(0.14)	(0.12)	(0.08)	(0.15)	(0.11)	(0.10)
Cons	211.56**	205.98***	199.26*	177.42**	209.61**	214.92***	227.04***	198.35**	187.85***	216.69***
	(84.97)	(59.08)	(102.46)	(80.41)	(70.77)	(60.57)	(41.88)	(75.21)	(55.32)	(47.27)
R^2	0.414	0.628	0.328	0.509	0.524	0.588	0.742	0.500	0.679	0.706
N	15	15	15	15	15	15	15	15	15	15

TIMSS 2003 mathematics cognitive scales and PISA 2003 mathematics subscales

TIMSS: (app) – applying; (kno) – knowing; (rea) – reasoning. **PISA:** see above

	m1	*m1*	*m1*	*m2*	*m2*	*m2*	*m3*	*m3*	*m3*	*m4*	*m4*	*m4*
	app	*kno*	*rea*	*app*	*kno*	*rea*	*app*	*kno*	*rea*	*app*	*kno*	*rea*
track	-18.91	-16.24	-12.84	-14.12	-11.26	-8.26	-27.97	-26.72	-23.53	-9.49	-7.34	-4.44
	(16.18)	(14.26)	(13.73)	(19.28)	(17.52)	(15.71)	(23.62)	(20.86)	(20.17)	(16.86)	(14.87)	(13.92)
TIMSS	0.64***	0.65***	0.69***	0.68***	0.69***	0.76***	0.59***	0.64***	0.69***	0.58***	0.60***	0.65***
	(0.13)	(0.11)	(0.11)	(0.15)	(0.14)	(0.13)	(0.19)	(0.16)	(0.17)	(0.14)	(0.12)	(0.12)
Cons	179.7**	175.8***	153.9**	161.1*	157.5**	122.9*	209.9**	187.3**	163.1*	207.4***	199.6***	175.2**
	(64.02)	(55.22)	(56.70)	(76.29)	(67.82)	(64.89)	(93.47)	(80.77)	(83.33)	(66.75)	(57.57)	(57.52)
R^2	0.620	0.697	0.714	0.559	0.627	0.694	0.361	0.489	0.513	0.550	0.641	0.680
N	15	15	15	15	15	15	15	15	15	15	15	15

* p<0.10, ** p<0.05, *** p<0.01

Table A5. DD with TIMSS subscales and PISA overall score in science and problem solving.

TIMSS 2003 and PISA 2003
TIMSS: life, physical, and earth sciences. **PISA:** science and problem solving overall scores

	Science Earth sciences	Science Life Sciences	Science Physics	Problem Earth sciences	Problem Life Sciences	Problem Physics
track	-6.27 (12.12)	-13.33 (11.09)	-2.45 (11.31)	-9.57 (17.80)	-18.69 (16.44)	-4.56 (17.59)
TIMSS	0.59*** (0.11)	0.52*** (0.08)	0.55*** (0.09)	0.75*** (0.15)	0.67*** (0.12)	0.68*** (0.14)
Cons	201.83*** (52.26)	237.94*** (39.77)	223.65*** (44.59)	125.77 (76.73)	168.83** (58.96)	160.03** (69.37)
R^2	0.681	0.743	0.718	0.607	0.677	0.609
N	15	15	15	15	15	15

Table A6. DD with TIMSS subscales in mathematics and PISA score in problem solving.

TIMSS 2003 subscales in mathematics and PISA 2003 problem solving score
TIMSS: all subscales in mathematics (see table A3). **PISA:** problem solving overall score

	Algebra	Data…	Fractions…	Geometry	Measurement	Applying	Knowing	Reasoning
Track	-20.21 (17.50)	-2.66 (13.03)	-23.14 (20.70)	-10.09 (14.14)	-20.18 (14.32)	-20.26 (18.90)	-17.53 (16.63)	-14.26 (14.98)
TIMSS	0.69*** (0.13)	0.65*** (0.09)	0.72*** (0.18)	0.75*** (0.11)	0.68*** (0.10)	0.70*** (0.15)	0.72*** (0.13)	0.78*** (0.12)
Cons	157.95** (65.92)	176.23*** (44.44)	145.59 (86.32)	122.41* (56.62)	162.20*** (50.09)	152.34* (74.80)	145.32** (64.37)	112.50* (61.87)
R^2	0.639	0.785	0.512	0.751	0.756	0.581	0.668	0.725
N	15	15	15	15	15	15	15	15

Table A7. Tracking effects in reading literacy by parental education in PIRLS and PISA.

PIRLS 2001 **PISA 2000**	*Tertiary*	*Upper/post secondary*	*Lower-secondary*	*Primary*	*Missing*
Track	-30.32** (14.41)	-38.22** (14.75)	-62.75*** (18.59)	-51.15*** (17.82)	-47.53*** (14.79)
PIRLS	0.62*** (0.20)	0.64*** (0.21)	0.70*** (0.22)	0.62*** (0.17)	0.60*** (0.17)
Constant	193.92* (107.51)	191.35* (104.27)	158.49 (99.84)	166.26** (75.99)	165.56* (80.92)
Adj. R-squared	0.341	0.354	0.449	0.505	0.480
N	22	22	22	22	22

PIRLS 2001 **PISA 2003**	*Tertiary*	*Upper/post secondary*	*Lower-secondary*	*Primary*	*Missing*
Track	-17.53* (9.63)	-29.95*** (9.95)	-49.05*** (14.29)	-18.98 (25.79)	-66.45*** (15.65)

PIRLS	0.39** (0.16)	0.52*** (0.15)	0.46** (0.17)	0.88** (0.30)	0.51** (0.20)
Constant	314.81*** (84.86)	246.33*** (74.16)	260.31*** (78.92)	42.09 (138.84)	203.63** (92.27)
Adj. R-squared	0.250	0.455	0.472	0.429	0.517
N	19	19	19	17	19

* p<0.10, ** p<0.05, *** p<0.01

Table A8. Tracking effects in science and mathematics by the number of books at home in TIMSS 2003 and PISA 2003.

Science	**How many books do you have at home?**					
	0-10	*11-25*	*26-100*	*101-200*	*More than 200*	*Missing*
Track	-14.43 (15.82)	-14.79 (15.44)	-8.18 (14.62)	-1.41 (13.64)	-3.51 (13.48)	-39.69** (16.73)
TIMSS	0.45*** (0.13)	0.59*** (0.14)	0.57*** (0.14)	0.48*** (0.12)	0.66*** (0.12)	0.53*** (0.13)
Constant	234.45*** (57.79)	178.81** (68.93)	198.57** (70.35)	262.55*** (60.78)	197.70*** (60.39)	205.56*** (58.91)
Adj. R-squared	0.421	0.510	0.517	0.520	0.685	0.529
N	15	15	15	15	15	15
Mathematics	**How many books do you have at home?**					
	0-10	11-25	26-100	101-200	More than 200	missing
Track	-29.24 (20.74)	-26.74 (19.46)	-19.07 (19.91)	-10.54 (18.00)	-9.03 (18.95)	-56.67** (20.96)
TIMSS	0.55*** (0.16)	0.70*** (0.17)	0.66*** (0.17)	0.61*** (0.15)	0.63*** (0.16)	0.56*** (0.17)
Constant	199.92** (71.58)	130.16 (79.12)	162.67* (86.80)	198.23** (78.47)	212.37** (81.84)	199.92** (72.94)
Adj. R-squared	0.413	0.530	0.470	0.507	0.505	0.465
N	15	15	15	15	15	15