Department of Economics and Statistics "Cognetti de Martiis"
Campus Luigi Einaudi, Lungo Dora Siena 100/A, 10153 Torino (Italy)
www.est.unito.it

Dipartimento
Economia e Statistica
Cognetti de Martiis

Working Paper Series

07/16

LEARNING INEQUALITIES BETWEEN PRIMARY AND SECONDARY SCHOOL. DIFFERENCE-IN-DIFFERENCE WITH INTERNATIONAL ASSESSMENTS

DALIT CONTINI and FEDERICA CUGNATA

UNIVERSITÀ DEGLI STUDI DI TORINO

The Department of Economics and Statistics "Cognetti de Martiis" publishes research papers authored by members and guests of the Department and of its research centers. ISSN: 2039-4004

Learning inequalities between primary and secondary school. Difference-in-difference with international assessments.

Dalit Contini*, Federica Cugnata**

* University of Torino
** University Vita-Salute San Raffaele

Abstract.

Evaluating the effect of institutional features by exploiting cross-country variability with cross-sectional data is difficult. Difference-in-difference strategies are sometimes employed to reach identification. In this paper, we discuss the difference-in-difference strategies adopted in the literature to evaluate the effect of early tracking on learning inequalities using surveys administered to children of different grades. In their seminal paper: "Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries" *Economic Journal* (2006), Hanushek, and Woessmann analyze the effect of early tracking on inequalities with two-step analysis. Other scholars, instead, focus on the social background regression coefficient, using individual-level models applied to pooled data from all countries. We demonstrate that since test scores are measured on different scales at different surveys, pooled data strategies may yield to completely uninformative results. Against this background, we use data on reading literacy in PIRLS 2006 and PISA 2012 and carry out two-step difference-in-difference analyses on the effect of early tracking on social background learning inequalities.

Keywords. Achievement inequalities, international assessments, early tracking**,** cross-sectional data, non-equated scores, difference-in-difference, pooled models, two-step estimation.

JEL classification: I24, C10

1. Introduction

The persistency of educational inequalities among different socioeconomic and demographic groups is an issue of major concern among social scientists. Along large differentials in educational attainment, the development of standardized learning assessments has highlighted the existence of substantial achievement inequalities among children of the same age or school grade in many countries. Since learning processes are cumulative (Cunha *et al*. 2006), the way inequalities evolve throughout childhood in different contexts is also of great interest. The ideal dataset to analyze the dynamics of learning inequalities is longitudinal, with achievement measured on the same scale at different age/grades (i.e. vertically equated scores). This would allow evaluating achievement growth for each child and relating this growth to individual, family and context factors, and to prior achievement. However, longitudinal data and equated scores are often unavailable. To describe the development of inequalities with cross-sectional data it is common practice to compare differentials between socio-demographic groups over assessments held at different school years; if test scores are non-equated, they are used in standardized form.

Parallel to national studies, the development of international surveys like PISA, TIMSS and PIRLS has revealed remarkable cross-country variability in children's competencies and in the extent to which ascribed individual characteristics affect learning (OECD, 2010a: OECD 2010b; Mullis *et al.* 2012; Mullis *et al.* 2012). Moreover, by exploiting the institutional variability existing at the cross-national level, these assessments also allow to analyze the role played by characteristics of educational systems (e.g. Hanushek and Woessmann, 2006; Ammermueller, 2007; Fuchs and Woessmann, 2007; Schuetz *et al.*, 2008). Since international surveys are cross-sectional, most of the existing literature focuses on achievement at a given age or school year.

Early tracking is indubitably the institution that has raised the greatest debate.[1] Arguments in favor of early tracking relate to the potential advantages of instruction with homogeneous groups of children. Opponents of early tracking argue that it fosters educational inequalities. Firstly, children of higher socioeconomic backgrounds, by receiving more familial support, tend to be more motivated and to perform better even at young age. Thus, early tracking exposes young children to homogeneous learning environments in terms of both ability and socioeconomic fabric. If peer effects operate, this segregation could go to the detriment of the children from disadvantaged backgrounds. Secondly, children of disadvantaged backgrounds are less likely to choose the academic track (and thus to be exposed to more ambitious learning content) even at similar levels of prior performance (Jackson, 2013). A strong influence of families on their offspring's educational choices – likely to enhance

[1] We refer to tracking policies where children following different educational programs are placed in different schools (as opposed to within-school ability streaming).

social origin inequalities, because costs and benefits may be evaluated differently across backgrounds and because of information asymmetries – is more likely to occur when tracking occurs at an early age, and with weaker ability restrictions (Checchi and Flabbi, 2007).

A number of studies analyze the effect of tracking on achievement inequality using single international assessment and estimate individual-level models on pooled data from all countries. By including an interaction term between family background and the system-level variable indexing tracking, they compare the family background coefficients between tracked and comprehensive systems. However, evaluating the impact of institutional features by exploiting cross-country variability is problematic with cross-sectional data, because it is difficult to control for unobserved system-level factors potentially affecting inequalities at all schooling stages. For this reason, some scholars propose to exploit two cross-sectional surveys held at different age or grades, and use difference-in-difference strategies. In their seminal work, Hanushek and Woessmann (2007) apply difference-in-difference to scores' dispersion, while Waldinger (2007), Jakuboski (2010), Van de Werfhost (2013), Ammermueller (2013) and Ruhose, Schwerdt (2015) apply difference-in-difference to family-background regression coefficients. Hanushek and Woessmann (2007) apply a two-step method: in the first step, they run separate models on individual data for each country; in the second step, they relate the estimates of the social background coefficient to country-level characteristics. The other scholars mentioned above, instead, pool together the data from all countries and assessments, and estimate individual-level models with individual-level and system-level explanatory variables. Notably, these scholars do not reach the same conclusions: most of them find that early tracking has a detrimental effect on equity, whereas Waldinger (2007) finds no negative effects.

In this contribution, we compare these estimation strategies – two-step and pooled individual models – in terms of their capacity to deliver meaningful findings. We go beyond the general limitations of cross-country studies in inferring "causal" effects of system-level features, and examine the specific restrictions imposed by the models adopted in the literature. We show that pooled individual models rely on unnecessary and often untenable constraints, and thus may yield to meaningless results. Restrictions are particularly severe because the test scores released by international assessments are measured on *different scales*. Two-step estimation, instead, always yields to interpretable findings.

The paper is organized as follows. In Section 2 we specify a simple achievement growth model, describe the mechanisms at play, relate them to the model's structural parameters, and discuss what information different types of data (longitudinal vs. cross-sectional) and measurement scales (equated vs. non-equated scores) deliver on these mechanisms. In Section 3 we describe the difference-in-difference strategies employed in the literature to evaluate the effects of institutional features on

achievement inequalities between children of different family backgrounds, and highlight the underlying assumptions and the conditions for consistent estimation of "causal" effects, ignoring scaling issues. We conclude our line of reasoning in Section 4, where we analyze the additional problems of these difference-in-difference strategies, arising when the dependent variable is measured on different scales over time, as occurs for international learning assessments.

Finally, by employing the data on reading literacy in PIRLS 2006 and PISA 2012, we carry out two-step difference-in-difference empirical analyses of the effect of tracking on learning inequalities. We replicate the analysis in Hanushek and Woessmann (2007) on overall inequality (as measured by the country standard deviation) with more recent data, and perform new analyses on social background inequalities (identified by the social background regression coefficient). Altogether, our results, summarized in Section 5, provide evidence that early tracking contributes to increasing both overall inequalities and the differentials among children of different social origin.

2. A simple achievement growth model.

Our starting point is a stylized model of achievement growth. In this section we present this model, and discuss the identification of the structural parameters with different types of data and different score measurements. We show that with cross-sectional data and achievement measured on different scales as children age – as occurs for international learning assessments – the relevant parameters are generally unidentified. More specifically, we show that the comparison of regression coefficients does not convey much information on the development of inequalities. This is a relevant point because, as we will see in section 3, what difference-in-difference strategies with pooled individual models do in essence, is comparing the difference between social background regression coefficients at different surveys across tracking regimes.

Consider a stylized model of learning development according to which abilities cumulate over time, so that achievement at time t equals achievement at time t-1 plus a growth component. This can be viewed as an ideal model of cognitive ability, assuming it can be measured on a meaningful interval scale and that it evolves linearly. Initial ability and growth may also be affected by individual ascribed characteristics such as gender and family background (e.g. socioeconomic status, minority, ethnic or immigrant origin). Children from advantaged backgrounds tend to perform better because they live in more stimulating environments and receive more parental support, and/or because, due to information asymmetries, they are more capable to acquire relevant information on the schooling system and choose better schools.

Assume we have two cross sectional surveys assessing students' learning at different stages of the educational career, t=1 and t=2. In order to keep the formalization as simple as possible, we posit no

measurement error, so that test scores are perfect measures of cognitive ability. Assume that achievement at different ages is measured on the same scale (i.e. test scores are "vertically equated")[2]. Let y_2 be the score at *t*=2 and $\tilde{y}_1$ the score at *t*=1. To simplify the exposition, we refer to a single explanatory variable *x* (social background, in our current example) and assume that:

$$\tilde{y}_{i1} = \mu_1 + \rho x_i + \varepsilon_{i1} \quad (1)$$

Scores at *t*=1 and *t*=2 are related by:

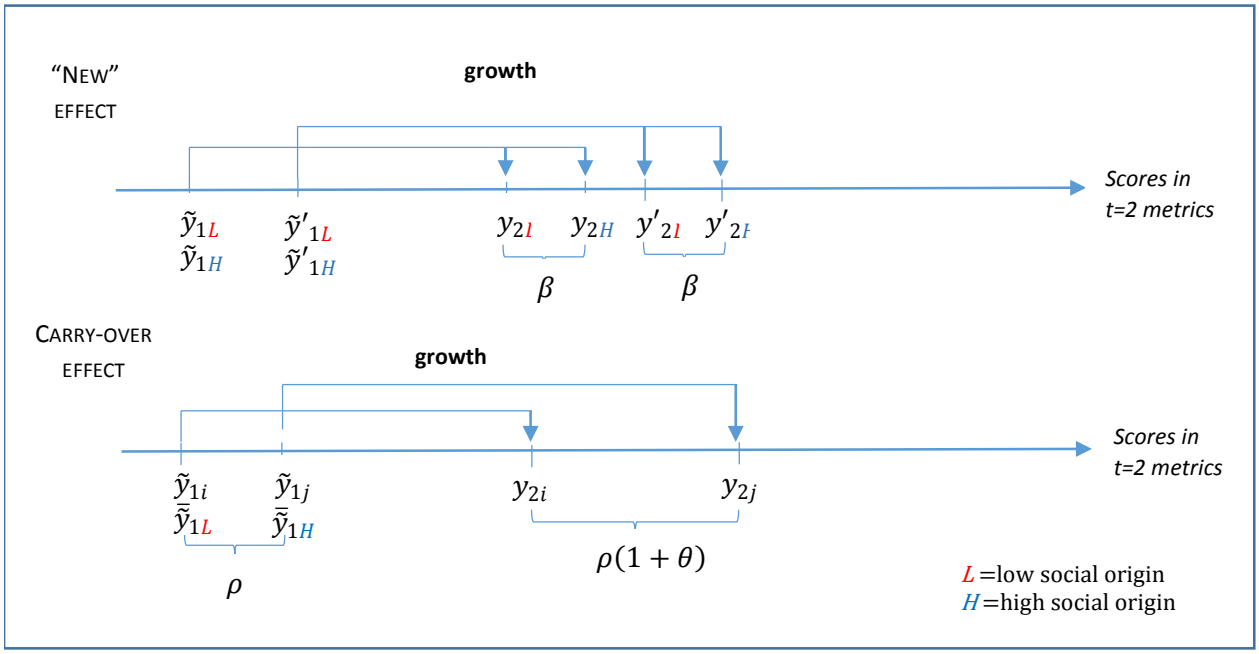
$$y_{i2} = \tilde{y}_{i1} + \delta_i \quad (2)$$

where δ_i is achievement growth. Growth is assumed to depend linearly on explanatory variables and may also depend on previous achievement:

$$\delta_i = \Delta + \beta x_i + \theta \tilde{y}_{i1} + \varepsilon_{i2} \quad (3)$$

β measures whether children of high backgrounds improve or worsen their performance between *t*=1 and *t*=2, relative to equally performing children of low backgrounds at *t*=1. We will refer to it as "new" *x*-effects. This is the most interesting mechanism because it involves social background directly. Instead, $(1+\theta)\rho$ are carry-over effects of pre-existing inequalities (see Figure 1).

Figure 1. The two mechanisms underlying the development of *x*-inequalities



[2] To create a vertical scale, scores from two tests at different age or grades are linked statistically through a process known as calibration, so that scores can be expressed on a common scale (Patz 2007). Practical and conceptual issues involved in vertical scaling (e.g. Bond and Lang 2012) are beyond the scope of this paper.

Ideally, we would like to estimate the average growth differential $E(\delta|x+1) - E(\delta|x) = \beta + \theta\rho$ and disentangle new *x*-effects and carryover effects, as they reflect substantially different mechanisms. Can we accomplish this with different types of data?

With *longitudinal data* and scores measured on a single scale as children grow older, the structural parameters are obviously identified. Assume now that achievement is measured on different scales. In this circumstance, $\tilde{y}_1$ represents the (unknown) score at *t*=1 in the measurement scale employed at *t*=2. Assume a linear relation between scales, so that $\tilde{y}_{i1} = \varphi + \omega y_{i1}$, where y_{i1} is the corresponding observed score. Note that φ and ω are not known and unidentifiable. The estimable model for y_{i1} is then:

$$y_{i1} = \frac{\tilde{y}_{i1} - \varphi}{\omega} = \frac{\mu_1 - \varphi}{\omega} + \frac{\rho}{\omega} x_i + \frac{\varepsilon_{i1}}{\omega} \tag{4}$$

while according to (1)-(3), the model for y_{i2} becomes:

$$\begin{aligned} y_{i2} &= \tilde{y}_{i1} + \delta_i = (\varphi + \omega y_{i1}) + \Delta + \beta x_i + \theta(\varphi + \omega y_{i1}) + \varepsilon_{i2} \\ &= \varphi(1+\theta) + \Delta + \omega(1+\theta) y_{i1} + \beta x_i + \varepsilon_{i2} \end{aligned} \tag{5}$$

By relating observed scores at two occasions, (5) has the structure of a panel data model with a lagged term. Note that $\omega(1+\theta)$ does not describe the dynamics of the learning process, as it depends on the unknown rescaling factor ω that allows to translate scores in the scale at *t*=1 into scores in the scale at *t*=2. Now, by conditioning on previous achievement, we can consistently estimate β. Instead, θ is unidentified: this means that we cannot measure absolute growth, nor test whether achievement of well performing children grows more or less than that of lower performing ones.

What can we infer on the development of inequalities with *cross-sectional* data? With simple substitutions, we obtain the cross-sectional model:

$$y_{i2} = \mu_2 + (\beta + (1+\theta)\rho) x_i + (1+\theta)\varepsilon_{i1} + \varepsilon_{i2} \tag{6}$$

The regression coefficient $\beta + (1+\theta)\rho$ represents the overall social background differential developed up to *t*=2 and is an estimable quantity with cross-sectional data.

Absolute scores

We may be tempted to evaluate whether in a given country social background inequalities have widened between two assessments, by comparing the average "growth" on observed scores:

$$(E[y_2|x+1] - E[y_1|x+1]) - (E[y_2|x] - E[y_1|x])$$

This amounts to evaluating the difference of regression coefficients at the two assessments shown in

(4) and (6). This difference is given by $\beta + (1+\theta)\rho - \frac{\rho}{\omega} = \beta + \theta\rho + (\omega - 1)\ \frac{\rho}{\omega}$. The term $\beta + \theta\rho$ is the overall true achievement growth differential: $E(\delta|x+1) - E(\delta|x)$, whereas $(\omega - 1)\ \frac{\rho}{\omega}$ has no substantive significance. Hence, with non-equated scores ($\omega \neq 1$) the difference between regression coefficients is meaningless.

Standardized scores

The most common strategy adopted in the existing literature to overcome the difficulties in comparing test scores measured on different scales is to standardize scores and compare average *z*-scores of individuals of different backgrounds as children age (e.g. Fryer and Levitt, 2004; Goodman *et al*., 2009; Reardon, 2011; Jerrim and Choi 2013). In a regression framework, this amounts to comparing *x*-coefficients from regressions on standardized scores. Results may be illustrated by simple graphs, and widening *z*-scores differentials across children's characteristics are interpreted as evidence of increasing inequalities.[3]

Indeed, differentials on standardized scores are invariant to the score metric. However, the sources of change remain unclear. The *z*-score differentials, obtained from (4) and (6), are:

$$E(z_1|x+1) - E(z_1|x) = \frac{\left(\frac{\rho}{\omega}\right)}{\sigma_{y_1}} = \frac{\rho}{\sigma_{\tilde{y}_1}}$$

$$E(z_2|x+1) - E(z_2|x) = \frac{(1+\theta)\rho+\beta}{\sigma_{y_2}} \qquad (7)$$

Hence:

$$[E(z_2|x+1) - E(z_2|x)] - [E(z_1|x+1) - E(z_1|x)] = \frac{(1+\theta)\rho+\beta}{\sigma_{y_2}} - \frac{\rho}{\sigma_{\tilde{y}_1}} \qquad (8)$$

Clearly, unless $\sigma_{y_2} = \sigma_{\tilde{y}_1}$, expression (8) does not allow identifying any of the relevant structural parameters. Notice also that we may observe negative (8) even if no genuine mechanism making a group catching up its previous disadvantage is at play. Consider $\beta = 0$ and $\theta = 0$. Due to (1) and (6), expression (8) becomes:

$$\frac{\rho}{\sqrt{\rho^2\sigma_x^2+\sigma_{\varepsilon_1}^2+\sigma_{\varepsilon_2}^2}} - \frac{\rho}{\sqrt{\rho^2\sigma_x^2+\sigma_{\varepsilon_1}^2}} < 0$$

Hence, in this case we observe a narrowing distance between children with different *x* simply because at *t*=2 there is higher variability.

What substantive mechanisms may make the score variance increase? Once again, assume we are interested in social background inequalities. To keep the notation simple, scores are depicted as being

[3] Similar graphs based on average percentiles are shown in Cunha *et al.* (2006) to provide a simple illustration of widening socioeconomic achievement gaps.

dependent on a single explanatory variable; however, other observed or unobserved factors may be involved. For example, gender, which, incidentally, is likely to be independent of social origin. Assume that for some reason gender inequalities widen between the two assessments, in the sense that at t=2 females will perform better on average than equally well performing boys at t=1. Although this mechanism should *not* affect the average social background growth differential in any way (neither directly nor indirectly, via previous performance), it will increase score variability at t=2, so that – in relative terms – high and low social background children eventually get closer. Therefore, even if in absolute terms the social origin scores differential does not change, it may decrease relative to the scores' standard deviation.[4]

It is important to notice that this is not necessarily a purely statistical artefact. On the contrary, it can be interpreted as a "real" effect, because the children of high and low social backgrounds do become more similar in some sense. Thus, if our aim is purely descriptive, comparing x-differentials of standardized scores at different occasions may make sense. However, the observed change could be due entirely to mechanisms that are totally unrelated with the grouping of interest: the use of standardized scores does not allow inferring the occurrence of any process making children with different x improving or worsening their performance relative to each other.

Another point is worth noticing. Assume that strong new x-inequalities develop between t=1 and t=2, so that β is positive and large. This will drive up the numerator of (7), i.e. $(1+\theta)\rho+\beta$, but it will also drive up the denominator, because it contributes to increasing the scores' variability. Since (7) is a growing but highly non-linear function of β, strong new x-inequalities will not necessarily be reflected in a large value of (8).

Summing up, the comparison of regression coefficients with cross-sectional data does not allow identifying any of the structural parameters of interest, neither using absolute scores, nor using standardized scores (Table 1). [5]

Table 1. Identifiable quantities with different data- and score-types

DATA TYPE	SCORES	INDIV GROWTH δ_i	OVERALL $\beta+\theta\rho$	"NEW" β	CARRY-OVER $\theta\rho$
Panel data	Same scale	YES	YES	YES	YES
	Different scales	NO	NO	YES	NO
Cross-sectional data (comparison of regression coefficients)	Same scale	NO	YES	NO	NO
	Different scale-absolute	NO	NO	NO	NO
	Different scale-standard.	NO	NO	NO	NO

[4] Expression (8) may be negative even if $\beta > 0$ and $\theta > 0$. However, a positive value of (8) implies a positive β (proof in Appendix A).

[5] We do not consider here more sophisticated pseudo-panel estimation strategies based on imputed regression, that under some conditions allow to estimate β with cross-sectional data (De Simone, 2013, Contini and Grand, 2015). As shown in Contini and Grand (2015) these strategies may deliver meaningful results only with very large samples, and hence are not appropriate for international learning surveys.

3. International assessments and the evaluation of early tracking

In this section, we review the empirical strategies most frequently adopted in the literature to analyze the effects of system-level features on achievement inequalities and compare the alternative difference-in-difference strategies in terms of underlying assumptions and restrictions.

By providing comparable measures of competencies across countries, international assessments are increasingly employed to analyze the effects of institutional features of educational systems (for an extensive review, see Hanushek and Woessmann, 2011). The most common modelling strategies are pooled-countries individual achievement models, with institutional features included as country-level explanatory variables, or two-step models – where the parameter of interest is estimated separately for each country in the first step, and its relation with system-level features is analyzed in the second. These contributions focus on the effects of institutions on mean performance (e.g. Woessmann 2005, Fuchs ad Woessmann, 2007, Woessmann 2010) or on equality of opportunity, usually operationalized as the socio-economic background gradient (Woessmann 2010).

The age of formal tracking into school-types offering substantially different educational programs varies greatly across countries: between age 10 in many German Laender to age 16 in UK and in Nordic European countries. Instead, the USA schooling system is comprehensive up to the end of secondary school, at age 18. To analyze the effect of the age of tracking on socioeconomic inequalities, Schuetz *et al.* (2008) and Brunello and Checchi (2007), examine single international assessments (TIMSS and IALSCC respectively) with a cross-sectional achievement model in which individual-level family background is interacted with a country-level dummy variable indexing early tracking. In essence, this amounts to comparing the family background coefficient between tracked and non-tracked countries. Interestingly, while Schuetz et al. (2008) find a substantive negative effect of tracking on children's performance at grade eight, Brunello and Checchi (2007) find the opposite effect on adult's cognitive skills.

Difference-in-difference strategies

The use of cross-sectional methods is open to criticism because they do not allow controlling for other cross-country institutional, cultural and societal differences affecting inequalities also before tracking takes place. To overcome this problem, in their seminal paper, Hanushek and Woessman (2007) analyze the variability of reading test scores (using the standard deviation and the distance between given percentiles) with a simple difference-in-difference strategy, exploiting two surveys held at different stages of the schooling career: PIRLS (4^{th} grade) and PISA (age 15). The idea is that in 4^{th} grade all children are in comprehensive school, whereas at age 15 in some countries students have already experienced educational tracking, in others they have not. Their empirical strategy basically

compares the change in scores' variability indexes occurring in this period, in countries with and without tracking. They find that in tracked systems variability increases over time relative to untracked ones, so they conclude that early tracking increases learning inequality.

Drawing on the work of Hanushek and Woessman (2007), a number of scholars (Waldinger 2007, Jakubowski 2010, Ammermuller, 2013; van de Werfhorst 2013) employ difference-in-difference strategies by estimating similar pooled-countries individual models, to analyze the effect of early tracking on achievement inequalities related to social origin, using the TIMSS assessment for math, or PIRLS and PISA for reading. Interestingly, these papers get to conflicting conclusions. Similarly, in a recent paper Ruhose and Schwerdt (2015) use difference-in-difference to study the effect of early tracking on achievement inequalities related to migrant background.[6]

We now examine these models more in detail. The simplest model is the one adopted by Waldinger (2007), Jakubowski (2010), Van de Werfhost (2013) and Ruhose, Schwerdt (2015):

$$Y_{ict} = \alpha_{0c} + \alpha_1 t + \xi_1 F_{ict} + \lambda_1 F_{ict} I_c + \xi_2 F_{ict} t + \lambda_2 F_{ict} t I_c + \varepsilon_{ict} \quad \text{(M1)}$$

where *F* is family background, *I* is the binary variable indexing early tracking, *t* is a binary variable indexing the secondary school survey, and subscripts *i*, *c* and *t* indicate the individual, country and survey. While the intercept is country-specific, all the other parameters are fixed, being allowed to vary only according to whether the system is tracked or untracked at age 15.[7] Net of other individual characteristics and school explanatory variables (not mentioned here for simplicity), the family background coefficient at *t*=1 is ξ_1 for untracked and $(\xi_1 + \lambda_1)$ for tracked countries, while at *t*=2 it is $(\xi_1 + \xi_2)$ for untracked and $(\xi_1 + \lambda_1 + \xi_2 + \lambda_2)$ for tracked countries.

The underlying assumptions of model M1 are very strong: (i) that family background inequalities at both surveys vary across countries only depending on tracking; (ii) that unobserved country characteristics may influence average scores, but do not affect family background inequalities.[8]

A more flexible model is estimated by Ammermuller (2013):

$$Y_{ict} = \alpha_{0ct} + \xi_{1c} F_{ict} + \xi_2 F_{ict} t + \lambda_2 F_{ict} t I_c + \varepsilon_{ict} \quad \text{(M2)}$$

---

[6] Despite their limited number, some of these studies are often cited in the literature. The strategy and findings of previous versions of Ammermueller (2013) and Waldinger (2007) are described in the influential Handbook of the Economics of Education (2011) in the chapters: Hanushek and Woessmann "The economics of international differences in educational achievement" (pg. 156), and Betts "The economics of tracking in education" (pg. 367).

[7] Some versions of this model allow for correlation between the errors terms within country clusters.

[8] Additional restrictions, involving also M2, are that the error term has the same variance across countries and that the coefficients of all other control variables, for example age of the child or gender, are fixed across surveys and countries. However, as shown by Guiso *et al.* (2008), gender inequalities greatly differ across countries. Limitations of pooled data models and their comparison with two-step estimation when individual variables vary across countries in cross-sectional studies are discussed in Heisig *et al.* (2015).

Here the intercept freely varies across countries and over time. Moreover, and this is the main point, the family background coefficient in primary school ξ_{1c} is unconstrained, while its variation between *t*=1 and *t*=2 depends only on tracking (the variation is ξ_2 for untracked countries and $\xi_2 + \lambda_2$ for tracked countries). Hence, the coefficients at *t*=2 are $(\xi_{1c} + \xi_2)$ for untracked and $(\xi_{1c} + \xi_2 + \lambda_2)$ for tracked countries.[9] The underlying assumptions are weaker than in model M1, because unobserved country characteristics are allowed to affect family background inequalities at *t*=1; instead, the *change* in family background inequalities between *t*=1 and *t*=2 is allowed to vary across countries only according to the tracking regime. Moreover, this change is fixed, and may not depend on inequalities at *t*=1.

In both M1 and M2, the parameter of main interest is λ_2, representing the effect of tracking on family background inequalities. This is the so-called "difference-in-difference" (*DID*). For M1, *DID* is the difference between the coefficients at *t*=2 for tracked and untracked countries, minus the corresponding difference at *t*=1:

$$\lambda_2 = \{(\xi_1 + \lambda_1 + \xi_2 + \lambda_2) - (\xi_1 + \xi_2)\} - \{(\xi_1 + \lambda_1) - \xi_1\}$$

For M2, *DID* can be conceived as the difference in the regression coefficients at *t*=2 between tracked and untracked systems, given the regression coefficient at *t*=1 (i.e. at a given level of previous inequality):

$$\lambda_2 = (\xi_{1c} + \xi_2 + \lambda_2) - (\xi_{1c} + \xi_2).$$

An even more flexible strategy would be applying difference-in-difference on social background inequalities with two-step modelling, similarly to what has been done by Hanushek and Woessman (2007) on test scores variability indexes. In a first step, country- and age-specific social background regression coefficients can be estimated with within-country models. In a second step, the estimated coefficients at *t*=2 can be related to early tracking controlling for inequality at *t*=1, by estimating a simple regression model or by graphical inspection. Despite this approach is deliberately exploratory, if within-country cross-sectional estimates are reliable estimates of the corresponding population parameters, in principle it could yield to valid causal inference. The critical assumption is that the change in social background inequalities between *t*=1 and *t*=2 only depends on tracking, or on unobserved country-level characteristics independent of tracking. Clearly, second step regression models run on a handful of countries suffer from small sample size. Yet, the same issue holds for pooled-country models on individual data, as the relevant sample size to the estimation of regression

[9] Ammermuller (2013) also analyses the effects of other institutional characteristics changing between primary and secondary school.

coefficients of country-level explanatory variables is the number of countries (Wooldridge, 2010; Bryan and Jenkins, 2016).[10]

Let us now give a closer look to the implications that each of the strategies discussed above have on the relationship between family background coefficients at *t*=1 and *t*=2. To the sake of generality, let us change notation and indicate these coefficients as γ_1 and γ_2:

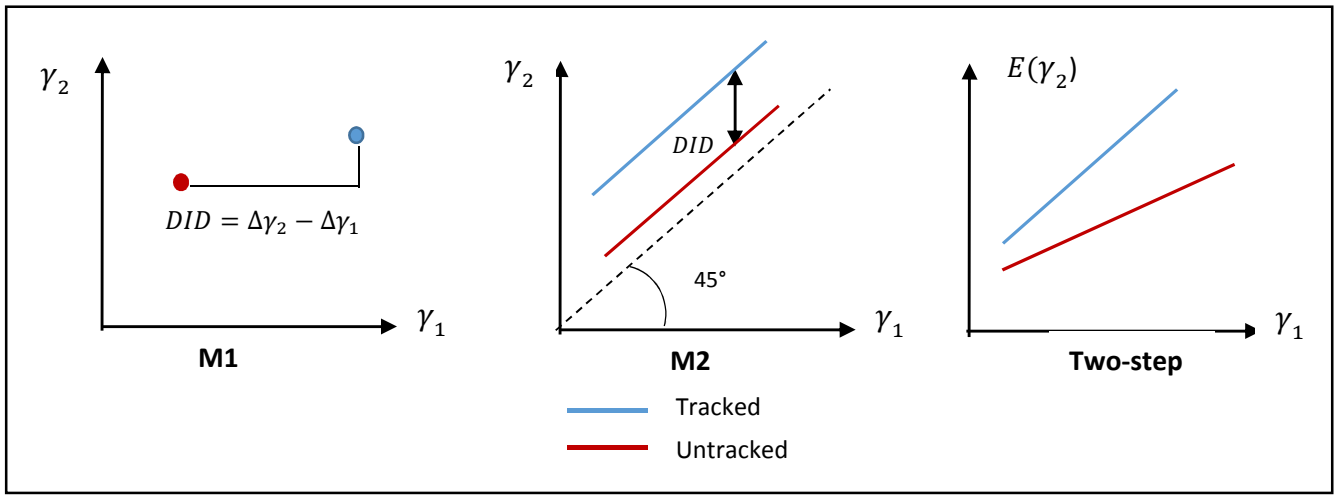
(i) In model M1 (Figure 2, left panel) γ_1 and γ_2 are constrained to be the same across countries, given tracking regime and survey, and *DID* amounts to $\Delta\gamma_2 - \Delta\gamma_1$, where Δ refers to the difference between tracked and untracked regimes.

(ii) In model M2 (Figure 2, central panel) γ_1 (in the previous notation ξ_{1c}) is allowed to vary freely across countries, while γ_2 (previously noted as $\xi_{1c}+\xi_2$ in untracked countries and $\xi_{1c}+\xi_2+\lambda_2$ in tracked countries) is constrained. Patently, the relation between these coefficients is: $\gamma_2 = \gamma_1 + \xi_2 + \lambda_2 I$. In a Cartesian coordinate system, this yields to parallel lines, and parallel to the first quadrant bisector.

(iii) In two-step modeling (Figure 2, right panel), there are no *a priori* constraints.

As we will show in the next section, these differences will turn out to be very important, not only from the perspective of better identification of the "causal" effect of early tracking (to control for unobserved country-level factors), but also in terms of the meaningfulness of the delivered results.

Figure 2. Difference-in-difference in pooled regression models and two-step analysis



NOTES. γ_1 = family background coefficient at *t*=1; γ_2 = family background coefficient at *t*=2.

[10] For an extensive comparison of pooled-country models and two-step estimation in a cross-sectional environment see Bryan and Jenkins (2016, supplementary material) and Heisig (2015).

4. Difference-in-difference with international scores

Let us briefly review the main points made in the previous sections. In section 2 we have specified a simple achievement growth model and shown that, when scores are non-equated, the difference of cross-sectional regression coefficients based on absolute scores is generally meaningless and the difference based on standardized scores conveys limited information on the development of inequalities as children grow older. In section 3 we have reviewed the difference-in-difference strategies employed in the literature and highlighted that, in essence, the individual models on pooled data identify the effect of early tracking on family background inequalities by comparing regression coefficients across surveys. In this section, we analyze how the results derived in the previous sections specifically apply to test scores delivered by international learning assessments, and derive implications on the validity of the alternative difference-in-difference strategies employed in the literature to evaluate the effects of early tracking on family background inequalities.

4.1 Test scores in international assessments

We start with the following question: should we conceive international test scores as absolute or standardized measures of achievement? International surveys rely on Item Response Theory (IRT) to produce measures of achievement. These methods take into account the items' difficulty, and in some cases the guessing probability and the items' discriminatory power.[11] Once IRT ability estimates are produced, they are standardized with respect to the mean and the SD of the *pooled* sample including all countries participating in the study. Transformed scores have mean 500 and SD 100.[12] In this sense, PISA, PIRLS and TIMSS produce standardized scores. Yet, a fundamental difference with the notion of standardized scores employed above is that in section 3 we were considering scores standardized *within* countries. Instead, international assessments use the same yardstick for all countries: if we compare two French individuals in PISA, we observe how many SD they are apart with respect to the cross-country SD, not to the French SD. To our end, this feature makes international scores alike absolute rather than standardized scores.

4.2 Difference-in-difference with original scores

Let us think of PIRLS (4^{th} grade) as the test at t=1 and PISA (age 15) as the test at t=2. As recalled above, international scores are standardized so to have overall mean 500 and average country SD

[11] In the IRT framework, the items' difficulty and individual ability are measured on the same scale. The ability of an individual is defined as the difficulty of the item for which the probability that the individual will provide a correct answer is equal to 0.50.

[12] Five random draws (the so-called "plausible values") from the posterior distribution of ability given the item's response pattern are taken for each individual.

100.[13] Measurement scales are non-equated. Thus, we may think of the original (equated) ability measures as being generated from models (1) and (6) – with as substituting ys. Roughly speaking, these ability measures are translated into international scores according to: $y_{it} = \left(\frac{a_{it} - \bar{a}_t}{\sigma_{a_t}(cc)}\right) 100 + 500$, where $\sigma_{a_t}(cc)$ is the average cross-country SD of the original measure at *t*.

Consistently with (4) and (6), for each country *c* international scores depend on the structural parameters ρ, β, θ introduced in section 2 according to:

$$y_{i1c} = \mu_{1c} + \rho_c F_1 x_{ic} + \varepsilon_{i1c} \quad (9)$$

$$y_{i2c} = \mu_{2c} + (\beta_c + (1 + \theta_c)\rho_c) F_2 x_{ic} + \varepsilon_{i2c} \quad (10)$$

with $F_1 = \frac{100}{\sigma_{\tilde{a}_1}(cc)}$ and $F_2 = \frac{100}{\sigma_{a_2}(cc)}$.

*Difference-in-difference with model M*1

In the most restrictive model M1, all the structural parameters of interest only depend on the tracking regime, so difference-in-difference (*DID*) amounts to:

$$DID = [(\beta_T + (1 + \theta_T)\rho_T) - (\beta_{\bar{T}} + (1 + \theta_{\bar{T}})\rho_{\bar{T}})] F_2 - (\rho_T - \rho_{\bar{T}}) F_1 \quad (11)$$

where T denotes tracked and $\bar{T}$ untracked educational systems at age 15. The building blocks of *DID* are regression coefficients expressed in each assessments' metrics, according to (9) and (10). Since $F_1 \neq F_2$, *DID* does not allow to identify any of the relevant mechanisms at play and this expression delivers meaningless results.

Still, it is important to recall that researchers do not only have information on *DID*, but also on the single regression coefficients at the two surveys. Hence, we may ask whether this empirical evidence taken as a whole allows to infer something meaningful on the sign of the relevant parameters β and θ. The general answer is no. In Table 2 we show how the empirical evidence on regression coefficients relates to *DID*, and what are the implications on the structural parameters. In the first three rows, depicting situations where the genuine effects of family background are smaller or equal in tracking countries than in non-tracking countries at *t*=1, and larger or equal at *t*=2, *DID* is necessarily positive, and the implications are that either β or θ (or both) must be larger in tracked than in untracked countries. In the last row, however, the family background coefficient is larger in tracked countries than in untracked countries, at both *t*=1 and *t*=2. Here *DID* could be either positive or negative. If *DID* is positive, once again either β or θ (or both) must be larger in tracked than in untracked

[13] We disregard here that the set of countries may vary across surveys, and assume that they remain the same over the two assessments.

countries. Instead, if *DID* is negative, nothing can be inferred.[14] Hence, we conclude that the difference-in-difference strategy based on M1 conveys little useful information on the relation between institutional features and the development of inequalities as children age.

Table 2. Empirical evidence and implications on structural parameters in model M1

Observed regression coefficient at *t*=1	Observed regression coefficient at *t*=2	Observed *DID*	IMPLICATIONS OF THE EMPIRICAL EVIDENCE
$\rho_T < \rho_{\bar{T}}$	$\beta_T + (1+\theta_T)\rho_T = \beta_{\bar{T}} + (1+\theta_{\bar{T}})\rho_{\bar{T}}$	Necessarily positive	$(\theta_T > \theta_{\bar{T}}) \cup (\beta_T > \beta_{\bar{T}})$
$\rho_T = \rho_{\bar{T}}$	$\beta_T + (1+\theta_T)\rho_T > \beta_{\bar{T}} + (1+\theta_{\bar{T}})\rho_{\bar{T}}$	Necessarily positive	$(\theta_T > \theta_{\bar{T}}) \cup (\beta_T > \beta_{\bar{T}})$
$\rho_T < \rho_{\bar{T}}$	$\beta_T + (1+\theta_T)\rho_T > \beta_{\bar{T}} + (1+\theta_{\bar{T}})\rho_{\bar{T}}$	Necessarily positive	$(\theta_T > \theta_{\bar{T}}) \cup (\beta_T > \beta_{\bar{T}})$
$\rho_T > \rho_{\bar{T}}$	$\beta_T + (1+\theta_T)\rho_T > \beta_{\bar{T}} + (1+\theta_{\bar{T}})\rho_{\bar{T}}$	Positive	$(\theta_T > \theta_{\bar{T}}) \cup (\beta_T > \beta_{\bar{T}})$
		Negative	NONE

Difference-in-difference with model M2

In the less restrictive model M2, inequalities at *t*=1 are unconstrained, thus ρ_c may freely vary across countries, regardless of the tracking regime. As shown in Section 3, *DID* represents the difference in the family background regression coefficients between tracked and untracked countries at *t*=2, given the coefficient at *t*=1 (see also Figure 2):

$$DID = (\beta_T + (1+\theta_T)\rho_c)F_2 - (\beta_{\bar{T}} + (1+\theta_{\bar{T}})\rho_c)F_2 = ((\beta_T + \theta_T\rho_c) - (\beta_{\bar{T}} + \theta_{\bar{T}}\rho_c))\,F_2 \quad (12)$$

In this case *DID* only depends on the score metric at *t*=2. This result is important because it implies that here *DID* is a meaningful quantity. A positive (negative) value of *DID* implies that the social background differential gap in achievement growth is larger (smaller) in tracked countries relative to untracked countries.

Nevertheless, this specification has still some undesirable constraints. In section 3, we derived that for M2 the relation between regression coefficients at the two assessments is: $\gamma_2 = \gamma_1 + \xi_2 + \lambda_2 I$. However, this constraint, implying a 45° degree line, is unnatural, and represents a threat to the validity of the results. To see this, recall the relation between regression coefficients and structural parameters: $\gamma_1 = \rho_c F_1$, $\gamma_2 = (\beta_T + (1+\theta_T)\rho_c)F_2$ in tracked countries and $\gamma_2 = (\beta_{\bar{T}} + (1+\theta_{\bar{T}})\rho_c)F_2$ in untracked countries. This implies that the general relation between γ_1 and γ_2 in each regime is linear and given by:

[14] The proof is trivial and not reported here. It is available from the author upon request. This result clearly applies in general when the regression coefficients differentials at *t*=1 and *t*=2 have the same sign and *DID* has the opposite sign.

$$\gamma_2 = \beta_T F_2 + (1 + \theta_T)(F_2/F_1)\gamma_1 \quad (13)$$

$$\gamma_2 = \beta_{\bar{T}} F_2 + (1 + \theta_{\bar{T}})(F_2/F_1)\gamma_1$$

Clearly, 45° lines cannot describe these relations. Moreover, if $\theta_T \neq \theta_{\bar{T}}$ the two resulting lines will have different slopes. Notice that the lines would not be parallel even without scaling issues (i.e. even if $F_1 = F_2$).

Difference-in-difference with two-step estimation strategy

In step 1, cross-sectional regression coefficients are estimated separately for each country at both surveys. Hence, there are no restrictions on any of the coefficients imposed. In step 2 we inspect the relation between the estimated social background coefficient at *t*=2 and institutional features, given the estimated social background coefficient at *t*=1, using either simple regression models with countries as statistical units, or graphical inspection. Difference-in-difference amounts to describing this relation, as depicted in Figure 2 (right panel).[15]

It is worthwhile noticing that, taking expressions (13) literally, the intercept gives information on new-inequalities β, and the slope gives information on carry-over effects θ. Thus, it would be possible to draw comparative conclusions on the structural parameters between tracked and untracked countries. However, since the intercept in quite instable with small sample size and the linear specification is only an approximation, we prefer to use a heuristic approach, and limit ourselves to analyzing the relation in an exploratory perspective.[16]

4.3 Difference-in-difference with within-country standardized scores

In Section 3 we have argued that comparing *x*-differentials of standardized scores at different occasions gives a descriptive picture of how *x*-differentials evolve in a given country, even if it does not provide information on the sources of the observed change. In this perspective, we may consider difference-in-difference on within-countries standardized scores.

Under the simplifying additional assumption that all model coefficients depend exclusively on the enforcing of tracking (so country SDs do not vary within a regime-type), standardized-*DID* for model M1 is equal to:

$$DID_{st} = \left(\frac{(\beta_T + (1+\theta_T)\rho_T)F_2}{\sigma_{y_2 T}} - \frac{((\beta_{\bar{T}} + (1+\theta_{\bar{T}})\rho_{\bar{T}}))F_2}{\sigma_{y_2 \bar{T}}}\right) - \left(\frac{\rho_T F_1}{\sigma_{y_1 T}} - \frac{\rho_{\bar{T}} F_1}{\sigma_{y_1 \bar{T}}}\right)$$

[15] Jakubowski (2010), Van de Werfhorst (2013) and Ammermueller (2013) argue that since the sample of countries differs across international survey and waves, difference-in-difference should be evaluated on rescaled scores (in order to obtain the same mean and standard deviation within the set of countries under study). The results presented in this section show instead that this strategy is unnecessary and does not solve any of the problems at stake.

[16] Since no countries display 0 inequality at *t*=1, the estimation of the intercept involves an extrapolation of the existing data; moreover, a small change in the slope's estimate may affect the intercept substantially.

The relevant parameters are clearly not identifiable. Yet, DID_{st} is not a meaningless quantity: by dividing by the country SD we obtain a metric-free measure, thus all its terms are comparable and in a purely descriptive perspective, some conclusions might still be drawn. Consider a positive DID_{st}. This result provides evidence that *for some reason*, the social background relative gap (relative to *each* country's SD) has increased more (or decreased less) in tracked countries than in untracked countries. As previously discussed, there are various possible reasons underlying this empirical result: different "new" social background effects given previous ability between tracking regimes, different carryover effects of prior ability, but also effects of other explanatory variables (even if independent of social background, as for example gender) that may influence the within-country scores' variability. In addition, since standardized regression coefficients may vary little even if β is large – because also σ_{y_2} will increase – a substantial increase in inequality in tracked regimes will not necessarily be reflected in a large value of DID_{st}.

5. Empirical analysis

5.1 Data and methods

We now carry out our own analysis on the effect of early tracking, exploiting the international surveys on reading literacy PIRLS 2006 and PISA 2012. PIRLS interviews children attending 4th grade (i.e. children at age 9-10), while PISA focuses on 15-year-old children. The time span between these surveys is approximately equal to the distance between age 9-10 and 15, so PIRLS 2006 and PISA 2012 can be thought as independent samples of a single birth cohort over time. We consider only Western world countries, as they share more similar schooling systems, societal organization and cultures, in order to reduce the risk of unobserved country level confounding factors. We select only those countries participating in both assessments, ending up with 24 countries (see Table 3). By tracking, we refer to the formal sorting process into educational programs with different academic content and learning targets, while we do not consider other forms of differentiation such as within-school ability-related streaming. We define countries as "tracked" if this sorting process on regular children takes place before age 15, as "untracked" otherwise. In our sample, we have 10 tracked and 14 untracked countries (Table 3).

In the empirical analyses, we focus on native children. The reason is twofold. Firstly, because we wish to avoid introducing an additional source of heterogeneity across-countries, due to the different composition of the immigrant background population in terms of countries of origin, immigration waves, socioeconomic fabric, and to the linguistic distance between countries of origin and destination. Secondly, because the relationship between social background and immigrant background educational inequalities is weak. Countries with low social background inequalities,

often display large immigrant background-specific penalties (i.e. controlling for social background, Borgna and Contini, 2014). In this light, analyzing only native children has the advantage of avoiding confounding effects of early tracking on social background inequalities due to the specific effects on the immigrant background population.

Table 3. Countries in the empirical analysis by tracking regime

TRACKED COUNTRIES AT AGE 15 (N=10)	UNTRACKED COUNTRIES AT AGE 15 (N=14)
Austria Belgium Bulgaria Germany Hungary Italy Luxembourg Netherlands Russian Fed. Slovakia	Canada Denmark France Israel Latvia Lithuania New Zealand Norway Poland Romania Slovenia Spain Sweden USA

In line with the methodological considerations developed in the previous sections, we apply two-step analysis. In the first step, we carry out within-country analyses and estimate the social background regression coefficients for each country. As indicators of social background, we include the log-number of books and a binary variable indicating whether at least one parent has tertiary education. In addition, we include gender and age as controls (see Appendix B for detailed definition on individual-level variables). In the second step, we analyze the relationship between estimated social background regression coefficients at t=2 and the tracking regime, given the social background regression coefficient estimates at t=1. We also perform a similar analysis on country test score standard deviations, as done by Hanushek and Woessmann (2007). Since these two measures (social background regression coefficients and standard deviations) convey different information, relating the two pictures allows getting a deeper understanding of the role of early tracking on the development of inequalities.

5.2 Empirical results

5.2.1 Difference-in-difference with pooled individual regression models

For illustrative purposes, at first we show the results of difference-in-difference estimation on pooled-countries individual models M1 and M2, with the tracking regime as the variable of main interest and

individual level characteristics as controls. In Table 4 we report the results on the coefficient of the interaction terms λ_2. In the second column we report the estimate of the corresponding coefficient for the log number of books, in the third the estimate of the corresponding coefficient for the variable indexing parental tertiary degree. In the last column, under the heading REG, we report the estimated linear combination of these two coefficients, allowing to highlight the effects of tracking on the differentials between children with tertiary educated parents and "many" books (500), and children with non-tertiary educated parents and "few" books (5) books, controlling for gender and age.

Table 4. Difference-in-difference results of pooled-countries regression

	ln(n°books) (1)	tertiary degree (2)	REG [ln(500)*(1)+(2)]- ln(5)*(1)
Model M1	-5.92**	0.84	-26.41*
Model M2	4.70*	1.57	23.24*
N individuals	240,271		
N countries	24		

NOTES *p-value<0.05, ** p-value<0.01, *** p-value<0.001

The results of the two models go in the opposite direction. According to M1, early tracking has a beneficial effect on social background inequalities; according to M2, early tracking contributes to increasing them. These conflicting results are due to the unnecessary restrictions, particularly severe in case of model M1 that, as shown above, may even produce meaningless results.

5.2.2 Two-step analysis: First step results

In the first step, we analyze data by country and survey. We compute descriptive statistics, including the standard deviations at *t*=1 and *t*=2, and estimate cross-sectional individual regression models with scores as the dependent variable, and socio-demographic characteristics as explanatory variables. The full set of first step results is available in Appendix C. In the following paragraph, we report some interesting correlations on absolute measures of inequality and country rankings.

Overall inequalities and social background inequalities

At each stage of the educational career, the total variance of test scores in each country can be decomposed into a component explained by social background and an unexplained component. More specifically, under the usual OLS assumptions: $\sigma_y^2 = \gamma^2\sigma_x^2 + \sigma_\varepsilon^2$. Hence, overall inequality depends on the social-background-specific effect (γ), on the variability of social background in the population (σ_x^2), and on the influence of other factors independent of social background (σ_ε^2). This simple relation clearly shows that overall achievement inequality and inequality between social backgrounds

are distinct phenomena: their relation is positive, but need not to be strong.[17] As shown in Table 5 (columns 3-4), the cross-country correlation between SD and REG (as defined in Table 4) is 0.617 at *t*=1 and 0.659 at *t*=2. If we consider country rankings instead of absolute values, we obtain 0.667 at *t*=1 and 0.578 at *t*=2.

Not surprisingly, countries displaying larger inequality in primary school also tend to display larger inequalities in secondary school (Table 5, columns 1-2). Correlations between social background differentials (REG) are stronger than between standard deviations, and substantially larger within tracked countries than within untracked countries. Interestingly, the correlation coefficient between ΔSD and ΔREG displayed in column 5 (where Δ refers to the difference between *t*=2 and *t*=1) computed on rankings is 0.738, i.e. positive and quite large (we do not compute the correlation on original scores, because, as we have seen, ΔREG has no substantive meaning). This tells us that countries raising their relative position with respect to overall inequality also tend to raise their relative position with respect to social background inequality.

Table 5. Cross-country correlations on absolute measures and rankings

ABSOLUTE MEASURES					
	(1)	(2)	(3)	(4)	(5)
	SD1, SD2	REG1, REG2	SD1, REG1	SD2, REG2	ΔSD, ΔREG
Tracked	0.699	0.816	0.741	0.764	-
Untracked	0.500	0.477	0.591	0.607	-
All	0.492	0.569	0.617	0.659	-
RANKINGS					
	SD1, SD2	REG1, REG2	SD1, REG1	SD2, REG2	ΔSD, ΔREG
Tracked	0.587	0.760	0.806	0.446	0.830
Untracked	0.274	0.751	0.584	0.642	0.618
All	0.323	0.688	0.667	0.578	0.738

5.2.3 Second step results

In the second step, we analyze country-level inequality measures by relating them to the tracking regime. Focusing on overall inequality, we find that on average the standard deviation at *t*=1 (PIRLS) is larger in untracked than in tracked countries, whereas the relation reverts at *t*=2 (PISA), where tracked countries display (slightly) larger values (Figure 6). The relation reverts also when looking at social background inequalities, as the average achievement gap between high and low strata (REG) is slightly larger in untracked countries at *t*=1, while at *t*=2 it becomes much larger in tracked countries. We obtain similar results if we look at country rankings.

[17] A related measure of inequality considered in PISA reports is the proportion of the variance of scores explained by social background, i.e. $\gamma^2\sigma_x^2/\sigma_y^2$.

Table 6. Country-level measures of inequality and rankings

	Original scores				Country rankings			
	SD1	SD2	REG1	REG2	SD1	SD2	REG1	REG2
Tracked	64.4 (9.1)	95.1 (8.7)	82.7 (20.8)	134.1 (24.7)	10.1 (7.4)	13.8 (6.3)	12.3 (7.9)	15.6 (6.8)
Untracked	71.3 (11.3)	92.9 (8.7)	84.0 (20.9)	117.3 (14.7)	14.2 (6.5)	11.6 (7.7)	12.6 (6.8)	10.3 (6.5)

NOTES. SD in parenthesis. Rank: 1=smallest, N=largest

We now describe the results of difference-in difference analyses. Following the results in section 4, we estimate a simple regression model relating the country-level measures of inequality at t=2 to tracking, given inequality at t=1. As remarked above, the statistical units in this step are countries, thus the estimation suffers from small sample size (but the same issue holds for pooled country models on individual data, because the relevant sample size for country-level explanatory variables is the number of countries). Sample size obviously influences the standard errors of the estimates, which tend to be large; hence, the results may not be statistically significant. However, as remarked by Borgna and Contini (2014), this should not be too much of an issue, because the countries analyzed are the countries we are interested in, and cannot be considered as a random sample drawn from a larger population. In this sense, all the analyses should be viewed as intrinsically descriptive, in the sense that they provide direct information on the population of interest, and statistical inference issues are not involved. Despite this caveat, we still report the usual results on standard errors and statistical significance.

In a first model, we force the two lines – relative to tracked and untracked systems – to be parallel; in the second, we add an interaction term allowing them to display different slopes. Results clearly indicate that early tracking is associated with an increase in inequalities (Table 7). Given inequality in primary school, the social background differential is on average 17.6 score units (0.176 standard deviations, according to the overall OECD distribution) higher in tracked than in untracked countries, whereas the standard deviation is on average 5.5 score units higher. Models with the interaction term, although not statistically significant for the SD, show that the slopes are larger in tracked systems, meaning that a unit increase in inequality at t=1 is associated with a larger increase in inequality at t=2 in tracked than in untracked countries.

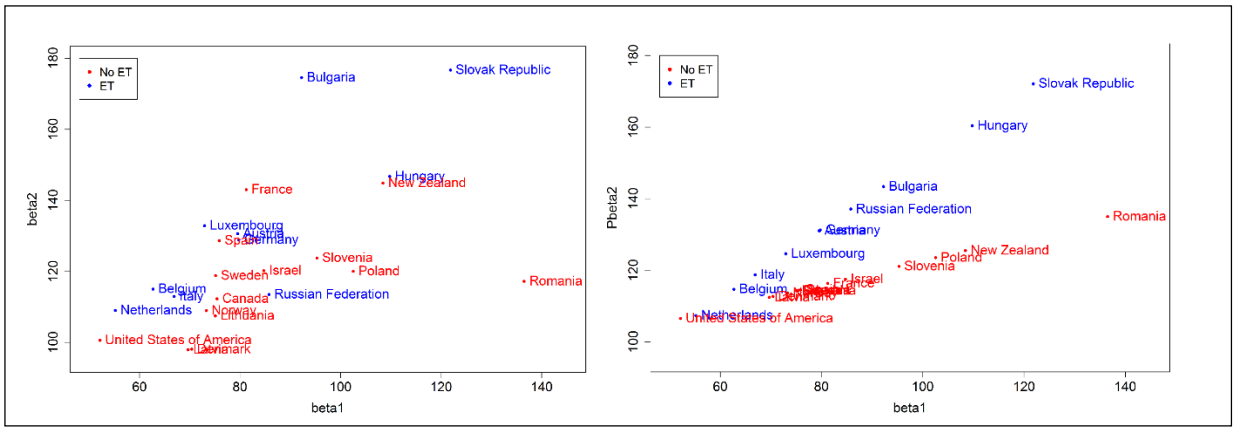
Table 7. Second step results. Cross-country regression models

INEQUALITY MEASURE (IM)				
	REG		SD	
	(1)	(2)	(1)	(2)
Constant	67.24***	88.95***	59.39***	65.63***
Track	17.59**	-35.18	5.45	-13.35
IM_1	0.596***	0.337*	0.471***	0.383**
Track* IM_1		0.634**		0.283
N	24	24	24	24
R^2	*0.504*	*0.597*	*0.334*	*0.359*

* p-value<0.10, ** p-value<0.05, *** p-value<0.01

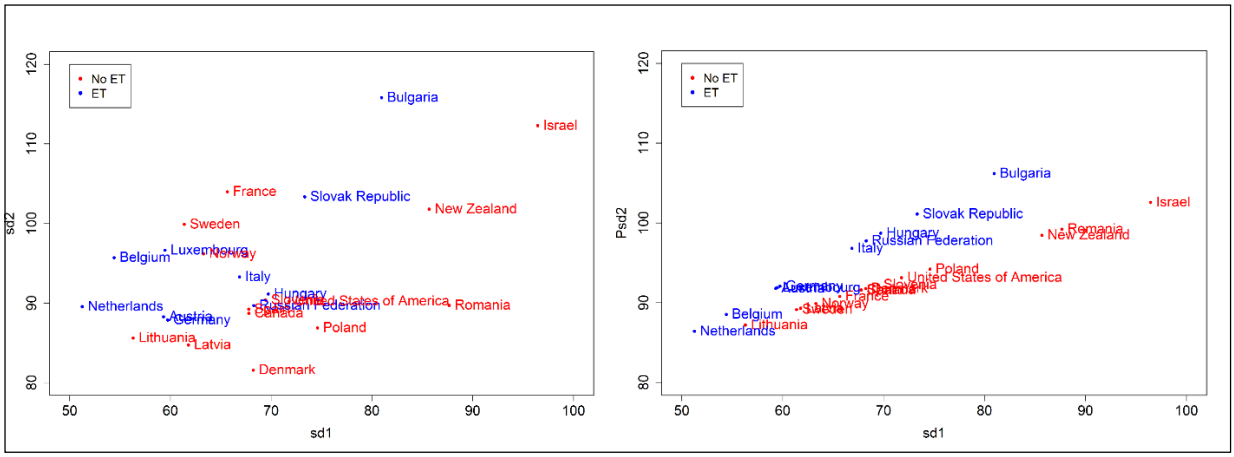
In Figures 3 and 4 we show the scatter diagrams depicting observed and predicted inequality measures in the two surveys. Firstly, these graphs show that in primary school both social background coefficients and standard deviations vary considerably across countries, but also within tracking regimes. Secondly, they allow us appreciating that at low levels of inequality in primary school there is little difference in (average) secondary school inequalities between countries with and without tracking, while at high levels of inequality in primary school, tracked systems become (on average) considerably more unequal. This pattern is more evident on the social background regression coefficient than on the standard deviation. A close inspection of Figure 3 also allows highlighting deviant cases (Russia among tracked systems and France among untracked systems) that could be the object of more in-depth qualitative analyses.

Figure 3. Observed and predicted social background differentials at *t*=2 given *t*=1



NOTE. Observed values in left panel. Predicted values in right panel

Figure 4. Observed and predicted test scores standard deviations at *t*=2 given *t*=1



NOTE. Observed values in left panel. Predicted values in right panel.

6. Concluding remarks

In this paper, we have carried out an in-depth discussion of difference-in-difference strategies exploiting international assessments administered to children of different age, to evaluate the effect on learning inequalities of institutional features varying over children's educational lives. In the existing literature, difference-in-difference has been carried out with two-step estimation by Hanushek and Woessmann (2007), who have analyzed overall inequalities (captured by test score variability measures), while other scholars have analyzed achievement inequalities related to family background (captured by the family background regression coefficients) with pooled-country individual level models. We have shown that while two-step analysis always delivers interpretable findings, pooled-country individual level models suffer from the imposition of restrictions hampering the meaningfulness of the results. The reason for the inappropriateness of most pooled-country individual level models, not acknowledged in the literature, rests on the fact that the dependent variable, test scores, are measured on different scales over time.

Our methodological discussion can be synthetized as follows. Firstly, starting from a stylized achievement growth model, we have derived the relation between the model's structural parameters and the difference between the cross-sectional regression coefficients of a given explanatory variable at different age of the children. If test scores are not vertically equated, this difference is shown to be meaningless when computed on original test scores, while it conveys limited information when computed on within-country standardized scores. Secondly, we have demonstrated that individual level models based on pooled data from all countries impose strong restrictions on the family background cross-sectional regression coefficients and on their relation. In the version of the most commonly used in the literature, *DID* is simply the difference between the cross-sectional regression

coefficients at the two surveys, but as we have shown, this quantity is essentially uninformative. Instead, difference-in-difference delivers interpretable results when performed with two-step estimation.

In the empirical part of the paper, we employ two-step estimation to analyze the relation between social background inequalities or overall inequalities and the tracking regime. Our findings point to a substantial increase of the social background coefficient in tracked relative to untracked countries at age 15, given inequality in primary school. Moreover, the gap increases with inequality in primary school: while at low levels we observe a small average difference between countries with and without tracking in secondary school, at high levels the difference is much larger. Results on standard deviations go in the same direction, but are somewhat weaker. In sum, early tracking appears to increase inequality, in particular by widening social background differentials.

A final remark on the limitations of our approach. As for all the empirical strategies exploiting cross-country variability in institutional features, the results are hardly interpretable in causal terms. The most important reason is that countries vary on a multitude of characteristics, so it is difficult to "hold other things constant". In this perspective, we consider two-step modeling as a more suitable approach than pooled modeling also because it clearly conveys the idea that we should consider our analyses as – highly informative – exploratory analyses. Using pooled-country models does not help in any respect; on the contrary, the formalization may give the impression to the naïve reader that the analyses are more rigorous. As we have seen, this is simply not true. A second problem is that the number of countries involved is typically small. Whatever the strategy, pooled models or two-step analysis, the relevant sample size for the estimation of the effects of country-level explanatory variables is the number of countries. Thus, we cannot rule out that the observed patterns are originated by random sources not under control. Finally, we have disregarded issues related to sampling variability in the first step. In principle, this could represent a problem, because it introduces measurement error in inequality measures in the second step estimation. However, with relatively large samples (and simple first-step models) this issue is unlikely to have a substantial impact on the results (Heisig *et al.*, 2015).[18]

[18] Second step models have been estimated also with procedures accounting for classical measurement error (an indication of the reliability is provided by the standard error of the estimates in the first step): the results on the effects of early tracking change very little.

References

Ammermueller, A. (2007) PISA: What makes the difference? Explaining the gap in test scores between Finland and Germany, *Empirical Economics*, **33**, 2, 263-287

Ammermueller, A. (2013) Institutional features of schooling systems and educational inequality: cross-country evidence from PIRLS and PISA, *German Economic Review*, **14**(2): 190-213

Betts J. R. (2011) The economics of tracking in education, in: *Handbook of the Economics of Education*, Vol. 3, edited by Hanushek E.A., Machin S., Woessmann L. Amsterdam: North Holland.

Borgna C., D. Contini (2014) Migrant achievement penalties in Western Europe. Do educational systems matter? *European Sociological Review*, 30, 5, 670-683

Brunello G., Checchi D. (2007) Does school tracking affect equality of opportunity? New international evidence, *Economic Policy*, **52**, 781-861

Bryan M.L., S.P. Jenkins (2016) Multilevel modelling of country effects: a cautionary tale, *European Sociological Review*, 32, 1, 3-22.

Bryan M.L., S.P. Jenkins (2016) Multilevel modelling of country effects: a cautionary tale, Supplementary Material, *European Sociological Review*, 32, 1.

Checchi D., L. Flabbi (2013) Intergenerational Mobility and Schooling Decisions in Germany and Italy: The Impact of Secondary School Tracks, *Rivista di Politica Economica* VII-IX (2013), 7-60.

Contini D., E. Grand (2015). On estimating achievement dynamic models from repeated cross-sections, *Sociological Methods and Research,* doi: 10.1177/0049124115613773.

Cunha, F., Heckman, J.J, Lochner, L. and Masterov, D.V. (2006) Interpreting the evidence on life cycle skill formation, in *Handbook of the Economics of Education* (edited by E. Hanushek and F. Welch), Chapter 12, pp. 697-812. Amsterdam: North Holland.

De Simone, G. (2013) Render into primary the things which are primary's. Inherited and fresh learning divides in Italian lower secondary education, *Economics of Education Review*, **35**, 12–23.

Fryer, R.G., S.D. Levitt (2004) Understanding the black-white test score gap in the first two years of school, *Review of Economics and Statistics*, 86, 2, 249, 281.

Fuchs, T., Woessmann, L. (2007) What accounts for international differences in student performance? A re-examination using PISA data, *Empirical Economics*, **32**, 2, 433-464

Goodman, A., Sibieta, L., Washbrook, E. (2009) Inequalities in educational outcomes among children aged 3 to 16. *Final report for the National Equality Panel*, UK

Guiso, L., Monte F., Sapienza P., Zingales L. (2008) Culture, gender and math, *Science* 30, 320-5880, 1164-1165.

Hanushek, E.A., Woessmann, L. (2006) Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries, *Economic Journal*, **116**, C63-C76.

Hanushek, E.A., Woessmann L. (2011). The Economics of International Differences in Educational Achievement. pp 89-200 in: *Handbook of the Economics of Education*, Vol. 3, edited by Hanushek E.A., Machin S., Woessmann L. Amsterdam: North Holland.

Heisig, J.P., Schaeffer M. and Giesecke J. (2015). Multilevel modeling when the effects of lower-level variables vary across clusters. A Monte-Carlo comparison of mixed-effects models,

cluster-rocust pooled OLS and two-step estimation. Unpublished manuscript.

Jackson M. (2013). *Determined to Succeed? Performance, Choice and Education*, Stanford University Press

Jakubowski, M. (2010) Institutional Tracking and Achievement Growth: Exploring Difference-in-Differences Approach to PIRLS, TIMSS, and PISA Data, in *Quality and Inequality of Education. Cross-National Perspectives* (eds J. Dronkers), pp 41-82. Springer.

Jerrim, J., Choi, A. (2013). The mathematics skills of school children: how does England compare to the high performing East Asian jurisdictions? *Working Paper of the Barcelona Institute of Economics* 2013/12

Mullis, I.V.S., Martin, M.O., Foy, P., Drucker, K.T. (2012). PIRLS 2011 International Results in reading. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

OECD (2010a) *PISA 2009 results: what students know and can do. Student performance in reading, mathematics and science*, Volume I.

OECD (2010b) *PISA 2009 results: overcoming social background. Equity in learning opportunities and outcomes*. Volume II.

Patz R. J. (2007) Vertical Scaling in Standards-Based Educational Assessment and Accountability Systems, CCSSO-Council of Chief State School Officers, Washington, DC

Reardon S.F. (2011) The widening academic achievement gap between the rich and the poor: new evidence and possible explanations, in Duncan G.J. and R.J. Murnane (eds) *Whither opportunity? Rising inequality, schools, and children's life chances*, Russel Sage Foundation.

Ruhose J., G. Schwerdt (2015) Does early educational tracking increase migrant-native achievement gaps? Difference-in-difference evidence across countries, IZA Discussion Paper 8903.

Schuetz, G., Ursprung, H.W., Woessman, L. (2008) Education policy and equality of opportunity, *Kyklos*, **61**(2), 279-308

Singer, J.D., Willett, J.B. (2003) *Applied longitudinal data analysis. Modelling change and event occurrence*. New York: Oxford University Press.

Van de Werfhost H. G. (2013) Educational tracking and social inequality in mathematics achievement in comparative perspective: two difference-in-difference designs. *Working Paper of the Amsterdam Centre for Inequality Studies*

Waldinger, F. (2007). Does ability tracking exacerbate the role of family background for students' test scores? *Working Paper of the London School of Economics*

Woessmann L. (2005). Educational production in Europe, *Economic Policy*, 20(43), 445-504

Woessmann L. (2010) Institutional determinants of school efficiency and equity: German states as a microcosm for OECD countries, *Jahrbücher für Nationalökonomie und Statistik*, 230(2), 234-270.

Wooldridge J.M. (2010) *Econometric Analysis of Cross-Section and Panel Data.* 2nd Edition. Cambridge MA: MIT Press.

Appendix A.

Proof that a positive difference of regression coefficients with standardized score implies $\beta>0$

Assume that (8) is positive:

$$\frac{(1+\theta)\rho+\beta}{\sigma_{y_2}} - \frac{\left(\frac{\rho}{\omega}\right)}{\sigma_{y_1}} > 0 \qquad \text{(A.1)}$$

As a consequence:

$$\beta > \left(\frac{\sigma_{y_2}}{\sigma_{y_1}}\frac{1}{\omega} - (1+\theta)\right)\rho$$

where $\frac{\sigma_{y_2}}{\omega\sigma_{y_1}} = \frac{\sigma_{y_2}}{\sigma_{\tilde{y}_1}}$.

From (1) and (6) we derive:

$$\sigma^2_{\tilde{y}_1} = \rho^2 var(x) + var(\varepsilon_1)$$

$$\sigma^2_{y_2} = [(1+\theta)\rho+\beta]^2 var(x) + (1+\theta)^2 var(\varepsilon_1) + var(\varepsilon_2)$$

The ratio is:

$$\frac{\sigma^2_{y_2}}{\sigma^2_{\tilde{y}_1}} = \frac{[(1+\theta)\rho+\beta]^2 var(x) + (1+\theta)^2 var(\varepsilon_1) + var(\varepsilon_2)}{\rho^2 var(x) + var(\varepsilon_1)}$$

$$= \frac{(1+\theta)^2\rho^2 var(x) + (1+\theta)^2 var(\varepsilon_1)}{\rho^2 var(x) + var(\varepsilon_1)} + \frac{[2(1+\theta)\rho+\beta^2]var(x)}{\rho^2 var(x) + var(\varepsilon_1)} + \frac{var(\varepsilon_2)}{\rho^2 var(x) + var(\varepsilon_1)}$$

$$= (1+\theta)^2 + \frac{[2(1+\theta)\rho+\beta^2]var(x)}{\rho^2 var(x) + var(\varepsilon_1)} + \frac{var(\varepsilon_2)}{\rho^2 var(x) + var(\varepsilon_1)}$$

Hence, for $\rho > 0$ and $\theta > -1$, we obtain $\sigma_{y_2}/\sigma_{\tilde{y}_1} > (1+\theta)$.

In conclusion, since these conditions are always satisfied, (A.1) implies $\beta > 0$.

Appendix B.

Table B.1 Variable definitions.

INDIVIDUAL VARIABLES	DEFINITION
POPULATION UNDER STUDY	
Natives	Children with at least one parent born in the country
SOCIAL BACKGROUND	
Books at home	Ln(n° books at home) Children report the number of books at home, based on pictures depicting different numbers of shelves. Classification in PIRLS is 0-10; 11-25; 26-100; 101-200, >200. Classification in PISA is 0-10; 11-25; 26-100; 101-200, 201-500, >500. The last two classes in PISA have been aggregated, so the two classifications are now identical. We have considered the central value in each class (500 in the highest class). In practice we use the following values: Ln(5)=1.61; Ln(13)=2.56; Ln(63)=4.14; Ln(150)=5.01; Ln(500)=6.21.
Parents with tertiary education	At least one parents with tertiary education=1 No parents with tertiary education=0
CONTROL VARIABLES	
Age	Country-specific quartiles' dummy variables (1°- 4°). We consider age in classes to allow for non-linear effects. The effect of age on test scores is unlikely to be linear. On the one side, the literature reports consistent evidence that older children tend to perform better (for example, in systems where regular children enter first grade in a given calendar year, children born in January tend to perform better than children born in December). On the other side, older children might be weaker. In some countries, there is flexibility in the age of first entry at school, so immature children might enter later, In other countries, poor performing children may be forced to repeat the school year, so older children are likely to be children who have experienced a grade failure. Quartiles are country-specific. This is particularly relevant for PIRLS, as regular age and age variability of 4th grade children varies substantially across countries (see Table B.2).
Gender	Female=0, Male=1

Table B.2 Age of native students in PIRLS (2006) and PISA (2012)

Country	PIRLS		PISA	
	Mean	SD	Mean	SD
Austria	10.31	0.423	15.81	0.292
Belgium	9.96	0.438	15.84	0.287
Bulgaria	10.87	0.478	15.80	0.280
Canada	9.92	0.350	15.84	0.283
Denmark	10.90	0.369	15.77	0.284
France	9.99	0.485	15.86	0.285
Germany	10.43	0.472	15.83	0.286
Hungary	10.65	0.467	15.73	0.288
Israel	10.07	0.356	15.69	0.284
Italy	9.68	0.318	15.76	0.285
Latvia	10.97	0.448	15.76	0.286
Lithuania	10.71	0.390	15.81	0.279
Luxembourg	11.34	0.514	15.82	0.291
Netherlands	10.21	0.467	15.70	0.286
New Zealand	10.03	0.329	15.76	0.286
Norway	9.79	0.289	15.79	0.291
Poland	9.89	0.302	15.71	0.279
Romania	10.92	0.488	15.72	0.274
Russian Federation	10.77	0.491	15.81	0.289
Slovakia	10.37	0.486	15.82	0.283
Slovenia	9.86	0.303	15.73	0.285
Spain	9.88	0.406	15.87	0.287
Sweden	10.85	0.315	15.73	0.278
United States of America	10.08	0.504	15.82	0.287

Appendix C.

First-step results

Table C.1 Scores standard deviation and standard error

Country	SD1	s.e	SD2	s.e
Austria	59.32	1.39	88.27	1.7
Belgium	54.42	0.89	95.7	1.78
Bulgaria	80.95	2.25	115.8	2.73
Canada	67.79	0.83	88.71	1.01
Denmark	68.22	1.28	81.56	1.75
France	65.65	1.01	103.98	2.33
Germany	59.73	1.23	87.86	1.78
Hungary	69.71	1.87	91.13	1.92
Israel	96.44	2.56	112.28	2.4
Italy	66.86	1.44	93.28	0.95
Latvia	61.78	1.45	84.73	1.83
Lithuania	56.30	1.26	85.60	1.5
Luxembourg	59.46	0.92	96.63	1.51
Netherlands	51.25	1.14	89.55	2.44
New Zealand	85.68	1.54	101.79	1.89
Norway	63.29	1.27	96.18	1.84
Poland	74.59	1.31	86.87	1.6
Romania	87.66	2.66	89.70	1.97
Russian Federation	68.27	2.15	89.68	1.57
Slovak Republic	73.32	2.19	103.35	3.16
Slovenia	69.44	0.98	90.37	0.9
Spain	67.77	1.3	89.22	1.13
Sweden	61.37	1.38	99.88	2.09
United States of America	71.78	1.43	90.22	1.76

NOTES. Native students. Explanatory variables: Gender (0=F, 1=M); Age in quartiles (ref cat=lowest quartile); ln(n° books); parent with tertiary education. Regressions estimates with own R routines (intsvy package) for plausible values and complex sampling, using student replicate weights.

Table C.2 First step results. PIRLS (2006)

COUNTRY	const.	gender	age_II	age_III	age_IV	ln(n°books)	tertiary	R2
Austria	504.41	-6.88	1.25	7.63	-14.49	10.85	29.54	12.88
se	6.38	2.69	3.29	3	4.33	1.32	2.72	1.57
Belgium	515.73	-6.08	-1.35	4.75	-20.73	8.13	25.24	14.45
se	5.06	2.48	2.71	2.97	3.22	0.94	2.19	1.48
Bulgaria	507.98	-17.52	-2.31	4.92	1.16	11.51	39.25	15.65
se	8.56	3.23	4.18	4.85	4.87	1.63	5.51	2.4
Canada	502.74	-10.85	6.7	11.49	-2.12	11	24.78	10.82
se	4.23	2.32	2.64	2.93	2.68	0.85	2.51	1.05
Denmark	500.71	-14.03	4.98	3.94	-5.28	11.08	19.39	8.78
se	7.26	3.38	4.1	4.82	5.19	1.21	3.82	1.49
France	482.73	-7.91	4.24	8.42	-25.84	10.51	32.84	18.9
se	5.24	2.96	3.4	3.14	3.96	1.09	3.03	1.51
Germany	502.74	-10.85	6.7	11.49	-2.12	11	24.78	10.82
se	4.23	2.32	2.64	2.93	2.68	0.85	2.51	1.05
Hungary	484.04	-2.47	2.86	6.61	-16.37	15.62	37.85	23.09
se	7.18	2.16	3.12	3.64	3.97	1.3	3.59	1.89
Israel	488.91	-13.03	15.51	15.33	23.19	6.01	57.06	14.43
se	11.07	5.2	5.97	6.47	7.1	2.41	4.69	2.22
Italy	513.62	-5.19	8.07	15.18	19.58	7.82	30.87	8.51
se	6.18	2.94	4	3.01	4.4	1.19	4.01	1.27
Latvia	511.49	-22.15	-3.87	0.2	-11.85	10.09	23.19	12.41
se	7.72	2.94	4.18	4.03	4.19	1.51	3.43	1.91
Luxembourg	519.59	-2.02	6.77	5.7	-36.91	12.52	15.29	18.27
se	5.23	2.31	2.6	3.02	3.76	0.92	2.9	1.62
Netherlands	526.37	-7.89	0.24	2.23	-19.76	7.29	21.62	14.54
se	5.31	2.11	3.16	2.72	3.78	1.26	3.19	1.95
New Zealand	466.74	-18.12	6.59	15.7	12.19	17.6	27.39	13.5
se	8.28	3.47	4.43	5.77	4.9	1.57	4.03	1.52
Norway	449.2	-16.5	4.4	10.89	14.53	10.27	26.05	13.49
se	6.56	3.31	3.68	4.08	4.98	1.3	3.46	1.74
Poland	463.23	-14.35	9.6	11.31	11.9	13.23	41.6	15.3
se	5	2.37	3.28	2.85	3.95	1.14	3.67	1.47
Romania	439.18	-14.73	3.12	3.23	-17.96	20.41	42.53	18.7
se	9.09	3.67	5.21	6.17	7.99	2.03	4.7	2.15
Russia	505.31	-14.42	5.9	13.83	3.66	12.61	27.71	14.65
se	8.64	2.9	3.51	3.35	3.8	1.53	3.66	1.84
Slovakia	452.18	-10.09	7.13	8.68	-10.47	19.46	32.22	21.26
se	7.07	2.39	3.13	3.41	4.97	1.6	2.9	2.21
Slovenia	473.39	-17.66	3.77	6.84	9.26	11.84	40.79	15.97
se	5.58	2.44	2.51	2.81	3.15	1.22	3.23	1.43
Spain	476.38	-0.83	3.53	9.71	-6.73	10.12	29.28	12.52
se	6.53	3.15	4.65	4.85	5.7	1.39	2.91	1.64
Sweden	500.43	-16.19	8.7	11.07	10.39	11.08	24.11	12.35
se	6.6	2.66	4.2	4.23	4.5	1.05	3.25	1.72
USA	508.54	-8.64	5.96	4.97	-15.73	11.32	n.a.	7.28
se	6.79	3.41	4	3.65	5.88	1.38	n.a.	1.29

NOTES. Within-country regressions. Native students. Explanatory variables: Gender (0=F, 1=M); Age in quartiles (ref cat=lowest quartile); ln(n° books); parent with tertiary education. Regressions estimates with own R routines (intsvy package) for plausible values and complex sampling, using student replicate weights.

Table C.3 First step results. PISA (2012)

COUNTRY	const.	gender	age_II	age_III	age_IV	ln(n°books)	Tertiary	R2
Austria	401.94	-30.09	3.49	6.41	5.41	24.02	19.96	22.92
se	5.92	4.49	4.51	4.63	5.22	1.17	3.62	1.71
Belgium	435.88	-28.42	7.44	13.66	16.01	20.68	19.80	16.94
se	5.72	3.22	3.09	3.40	3.31	0.97	3.09	1.12
Bulgaria	348.37	-59.75	0.74	0.87	4.13	30.72	33.09	32.06
se	7.21	4.03	4.52	4.42	4.43	1.56	3.67	1.73
Canada	438.98	-30.65	0.26	7.74	7.02	20.21	19.22	17.14
se	4.31	2.05	2.94	2.88	2.85	0.80	2.03	0.93
Denmark	434.17	-26.64	5.58	4.54	7.13	17.89	15.69	15.56
se	5.71	2.63	3.89	3.67	3.59	0.98	3.54	1.34
France	411.87	-36.89	0.82	10.41	13.23	28.53	11.60	24.29
se	7.07	3.35	4.14	4.13	4.96	1.51	4.18	1.6
Germany	422.13	-37.62	-1.76	7.52	10.89	24.15	17.65	24.69
se	6.9	2.70	3.45	4.05	4.57	1.35	3.35	1.53
Hungary	368.2	-34.52	6.68	9.28	12.7	27.97	17.90	30.56
se	6.45	3.53	4.18	4.13	4.71	1.14	3.84	1.9
Israel	418.29	-42.8	7.87	13.81	14.97	12.76	61.42	16.38
se	10.53	6.82	5.71	6.27	5.57	2.07	5.2	1.54
Italy	412.02	-34.1	-0.36	7.82	11.21	22.87	7.58	17.59
se	4.12	2.17	2.10	1.92	2.38	0.71	1.96	0.73
Latvia	431.1	-50.95	9.48	10.84	14.72	16.32	22.79	21.38
se	6.30	4.15	3.94	4.29	4.49	1.19	3.66	1.69
Luxembourg	378.69	-23.97	9.98	12.84	13.01	27.22	7.55	17.67
se	7.66	2.98	4.88	4.17	5.53	1.43	3.89	1.52
Netherlands	434.12	-22.58	1.7	5.45	9.44	22.69	4.51	18.23
se	5.87	3.02	3.61	3.96	3.81	1.20	4.93	1.62
New Zealand	405.21	-28.86	11.48	6.48	21.53	25.61	26.92	20.06
se	9.23	5.01	4.30	4.56	4.78	1.60	4.1	1.78
Norway	420.11	-37.57	5.16	14.83	11.41	22.69	4.48	16.65
se	7.30	3.22	4.09	4.28	4.73	1.27	4.02	1.23
Poland	449.34	-36.23	-1.25	6.54	4.75	18.78	33.51	21.47
se	5.99	2.67	3.13	3.98	3.66	1.26	3.37	1.47
Romania	375.75	-37.14	-5.31	2.13	-1.99	21.01	20.45	20.62
se	6.49	3.48	3.49	3.46	3.55	1.36	3.83	1.88
Russia	405.16	-35.09	5.38	8.54	5.96	16.11	39.32	18.3
se	6.57	2.97	3.35	3.57	3.95	1.17	3.73	1.6
Slovakia	343.4	-34.31	13.42	8.58	11.27	32.72	26.00	30.14
se	10.61	4.10	5.44	6.03	4.81	2.00	4.37	1.98
Slovenia	420.6	-47.44	-0.64	-0.74	8.11	20.44	29.66	23.83
se	4.69	2.78	3.71	3.60	4.01	1.06	3.05	1.12
Spain	393.83	-25.02	3.72	8.61	7.38	22.67	24.28	19.12
se	5.03	2.18	2.6	2.49	2.63	0.89	2.31	1.07
Sweden	397.59	-41.77	6.03	12.65	16.21	23.96	8.43	18.12
se	8.13	3.93	4.53	3.90	4.45	1.46	3.32	1.31
USA	424.35	-26.14	1.36	8.71	12.46	21.85	n.i.	16.70
se	6.43	3.05	3.97	3.56	3.99	1.53	n.i.	1.79

NOTES. Within-country regressions. Native students. Explanatory variables: Gender (0=F, 1=M); Age in quartiles (ref cat=lowest quartile); ln(n° books); parent with tertiary education. Regressions estimates with own R routines (intsvy package) for plausible values and complex sampling, using student replicate weights.

Table C.4 Social background differentials and standard error

	REG1	se	REG2	se
Austria	79.53	10.53	130.59	5.92
Belgium	62.70	7.42	115.03	5.11
Bulgaria	92.25	7.78	174.56	8.57
Canada	75.44	6.20	112.28	4.08
Denmark	70.40	8.97	98.07	5.73
France	81.24	7.57	143.00	7.67
Germany	79.76	8.51	128.87	6.27
Hungary	109.77	2.33	146.72	6.75
Israel	84.72	11.59	120.19	10.11
Italy	66.87	6.05	112.92	3.71
Latvia	69.66	7.14	97.95	6.28
Lithuania	75.08	7.09	107.48	4.62
Luxembourg	72.94	3.26	132.89	6.27
Netherlands	55.17	3.10	109.00	6.86
New Zealand`	108.43	7.80	144.85	8.42
Norway	73.32	3.41	108.97	6.89
Poland	102.53	10.12	119.99	6.61
Romania	136.53	20.41	117.19	7.80
Russian Federation`	85.76	5.63	113.50	7.10
Slovak Republic	121.84	3.66	176.67	10.03
Slovenia	95.30	4.28	123.77	4.60
Spain	75.87	3.44	128.66	4.72
Sweden	75.14	2.85	118.79	6.99
United States of America	52.12	7.41	100.61	7.05

NOTES. Under the heading REG we report test scores estimated differentials between children with tertiary educated parents and log n° books=6.21 (corresponding to 500 books), and children with non-tertiary educated parents and log n° books=1.61 (corresponding to 5 books), controlling for gender and age (see Tables E.2 and E.3). REG1 are estimates from PIRLS (2006). REG2 are estimates from PISA (2012). Standard errors of the linear combination obtained with own R routines (intsvy package) for plausible values and complex sampling, using student replicate weights.