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

Working Paper Series

17/18

HOW DO INSTITUTIONS AFFECT LEARNING INEQUALITIES? REVISITING DIFFERENCE-IN-DIFFERENCE MODELS WITH INTERNATIONAL ASSESSMENTS

DALIT CONTINI AND FEDERICA CUGNATA

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

How do institutions affect learning inequalities? Revisiting difference-in-difference models with international assessments

Dalit Contini*, Federica Cugnata**

* University of Torino
dalit.contini@unito.it

** Vita-Salute San Raffaele University

Abstract.

In this contribution, we discuss the difference-in-difference strategies employed in the literature to evaluate the effect of institutional features on learning inequalities exploiting international assessments administered at different age/grades. In their seminal paper, Hanushek and Woessmann (2006) analyze with two-step estimation the effect of early tracking on overall inequalities, measured by variability indexes. Later work of other scholars focuses instead on inequalities among children of different family backgrounds, using individual-level models on data pooled from different countries and assessments. We demonstrate that since test-scores are measured with different scales at different assessments, pooled individual models may deliver severely biased results. Instead, the scaling problem does not affect the two-step approach. For this reason, we advocate the use of two-step estimation also to analyze *family-background* achievement inequalities. Against this background, using PIRLS-2006 and PISA-2012 we conduct two-step difference-in-difference analyses, finding new evidence that early tracking fosters both overall inequalities and family background differentials in reading literacy.

Keywords: test scores, achievement inequalities, early tracking, international assessments, vertical scaling, cross-country analyses

(Revised version, October 2018)

1. Introduction

In spite of the fundamental principle that all children should have the same learning opportunities, large differentials are observed among socioeconomic and demographic groups in the share of students attending academic upper secondary programs and obtaining tertiary education (Jackson, 2013). Along inequalities in educational attainment, national and international standardized learning assessments have highlighted the existence of substantial differentials across social groups also in the children's level of competences and curricular knowledge at earlier stages of schooling. The persistency of educational inequalities is an issue of major concern among social scientists, both as a problem of social justice *per se*, and for its societal and economic consequences. In fact, the literature emphasizes education as one of the major factors affecting the degree of income inequality (De Gregorio and Lee, 2002) and social cohesion (Green, Preston and Janmaat, 2006), and there is ample evidence that the cognitive skills of the population and their distribution strongly affect economic growth (Hanushek and Woessmann, 2015).

The development of international surveys on children's learning like PISA, PIRLS and TIMSS – delivering comparable achievement measures across educational systems – has revealed large cross-country variability in average performance and in the degree of inequality across social groups. A key question is whether and how institutional differences affect the level and distribution of educational outcomes. By exploiting the institutional variability existing at the cross-national level, international assessments allow to investigate empirically the role played by the characteristics of school systems (for extensive reviews, see Hanushek and Woessmann 2011 and Woessmann 2016).

The age of tracking is indubitably the institutional feature that has raised the greatest debate. Tracking occurs when children choose between (or are placed into) different school-types to follow educational programs with different prestige level and learning targets. The age of formal tracking varies greatly across countries: between age 10 in many German states to age 16 in UK and in Nordic European countries. Instead, the American and Canadian schooling systems are comprehensive up to the end of secondary school, at age 18. 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 a 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 be detrimental to 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 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, 2013; Contini and Scagni 2011).

Because of its relevance, many scholars have analyzed the effect of tracking on achievement. Some studies exploit educational reforms put into effect in some regions or countries (Meghir and Palme, 2005 on Sweden; Malamud *et al*, 2011 on Romania; Piopiunik 2014 on Bavaria; Kerr *et al*. 2017 on Finland). However, specific institutional reforms are implemented only in few countries and typically at once, so the impact of institutions cannot always be investigated in this way. Moreover, one should rely on before and after comparisons that may confound the effects of policies with other country and cohort effects (Brunello and Checchi, 2007); even when they have high internal validity, the findings may not be easily generalized to different contexts.

Other studies exploit the cross-country institutional variability and utilize the international learning assessments to estimate educational production functions, i.e. individual-level models of achievement, on data pooled together from all countries. A number of contributions focus on the effect of tracking on family background inequalities at given age or stages of schooling (e.g. Brunello and Checchi 2007, Schuetz *et al*. 2008, Horn 2009, Woessmann 2010, Bol *et al.* 2014, Chmielewski and Reardon 2016). However, evaluating the impact of institutions exploiting cross-country variability is problematic with cross-sectional data, because of the difficulty to control for unobserved system-level factors potentially affecting inequalities at all schooling stages. For this reason, in their seminal work Hanushek and Woessmann (2006) propose to use two cross-sectional surveys held at different age or grades and employ difference-in-difference strategies. In particular, they apply difference-in-difference to test scores' *variability indexes*, finding that variability increases in early tracking relative to late tracking countries. More recently, other scholars have adapted their approach to analyze how early tracking affects learning inequalities across social groups by applying difference-in-difference to *family-background* differentials (Waldinger 2007, Jakuboski 2010, Van de Werfhost 2013, Ammermueller 2013, Ruhose, Schwerdt 2016). Hanushek and Woessmann (2006) use two-step estimation: in the first step, they estimate the variability indexes for each country and survey; in the second step, they relate these estimates to the early tracking indicator. The other studies, instead, pool together the data from all countries and assessments, and estimate individual-level achievement models with individual- and system-level explanatory variables.

The comparison of the behavior of the estimates in individual pooled-data models and two-step strategies in standard cross-sectional studies has been the object of recent methodological work (Heisig *et al.* 2017, Bryan and Jenkins 2016). In this paper, we analyze these strategies when applied

to difference-in-difference modeling. Our aim is to compare two-step and pooled individual models in terms of their capacity to deliver meaningful findings on the effect of institutional features on family-background achievement inequalities. More specifically, we address an issue that to our knowledge is completely missing in the sociology and economics of education literatures, related to the fact that test scores released by different international assessments are not vertically equated, i.e. achievement is *not measured on the same scale* as children grow up. We demonstrate that when the dependent variable follows different metrics over time, difference-in-difference estimation on pooled individual models relies on unnecessary and often untenable constraints, and thus may yield to meaningless findings. Instead, we show that this issue does not affect the two-step estimation strategy.

Against this background, by employing the data on reading literacy in PIRLS 2006 and PISA 2012, we carry out an empirical analysis of the effect of tracking on learning inequalities in reading literacy, using two-step analysis. Firstly, we replicate the analysis proposed by Hanushek and Woessman (2006) on the test score's standard deviation with more recent data; secondly, we analyze how tracking affects inequalities among children of different socioeconomic origin. Altogether, we provide new evidence that early tracking contributes to increasing overall variability and in particular the gap between children of different social backgrounds.

The remainder of the paper is organized as follows. In Section 2, we describe the difference-in-difference strategies employed in the existing literature to evaluate institutional effects on achievement inequalities. We start by describing the two-step approach employed by Hanushek and Woessamn (2006) to analyze the effect of early tracking on country-level variability measures, and then move to the individual pooled models used to study institutional effects on family background learning inequalities. We show that individual pooled models are quite restrictive and that in essence they estimate the effect of tracking by double differentiating the estimated (cross-sectional) family background regression coefficients between tracking regimes and learning assessments. Extending the approach of Hanushek and Woessamn (2006), we then propose a more flexible two-level model describing individual achievement within countries and then relating family-background regression coefficients to institutional variables. In Section 3, we address the scaling issue: starting from a simple learning growth model, we outline the mechanisms at play and show that if test scores at different surveys are not measured on the same scale – as occurs for international learning assessments – differentiating cross sectional regression coefficients conveys little information on how inequalities develop as children grow older. In Section 4 we analyze how the scaling issue affects the results of difference-in-difference models and demonstrate that the estimates of institutional effects delivered by pooled individual models may be severely biased. In Section 5, we describe our empirical analysis and discuss the results. Conclusions follow.

2. International assessments and the evaluation of institutional effects

International learning surveys were designed to evaluate education systems by testing the skills and knowledge of students of different age in different domains. The *Programme for International Student Assessment* (PISA) evaluates reading literacy, mathematics and science on children of age 15 (OECD 2014). The *Progress in International Reading Literacy Study* (PIRLS) focuses on pupils in grade 4 (Mullis *et al.* 2012a) and the *Trends in Mathematics and Science Study* (TIMSS) on pupils in grades 4 and 8 (Mullis *et al.* 2012b). By providing comparable measures of competencies across countries, these international learning surveys are increasingly employed to analyze how educational systems affect achievement (Hanushek and Woessmann, 2011, Woessmann 2016). In this section, we analyze the empirical strategies most frequently adopted in the literature to evaluate the effects of system-level features on achievement inequalities and compare difference-in-difference strategies in terms of their underlying assumptions and restrictions.

A number of contributions analyze test scores delivered by a *single* assessment administered at a given age or stage of schooling. While some studies focus on the effects of educational institutions (e.g. tracking, central examinations, school autonomy) on mean performance (e.g. Woessmann 2005, Fuchs and Woessmann, 2007, Woessmann 2010), others analyze the effects on inequality of opportunity, operationalized as family-background performance differentials (Brunello and Checchi 2007, Schuetz *et al*. 2008, Horn 2009, Woessmann 2010, Bol *et al.* 2014, Chmielewski and Reardon 2016). Focusing on the effect of early tracking, Schuetz *et al*. (2008) and Horn (2009) report a substantive negative effect of tracking on social background inequalities in children's performance, whereas Brunello and Checchi (2007) find the opposite effect on adult's cognitive skills. Bol *et al.* (2014) investigate how central examinations affect the association between tracking and family background inequalities. Chmielewski and Reardon (2016) provide evidence that tracking also enhances income achievement inequalities. A two-step approach is employed in some cases (Schuetz *et al*. 2008, Woessmann 2010, Chmielewski and Reardon 2016). In the first step, the parameter of interest is estimated separately for each country with individual-level achievement models, in the second, the relation between this parameter and system-level features is analyzed with a simple country-level model. Other scholars, instead, pool together the international data and estimate individual achievement models with institutional features as country-level explanatory variables (Woessmann 2005, Fuchs and Woessmann, 2007, Schuetz *et al*. 2008, Bol *et al.* 2014). Models focusing on inequalities also include an interaction term between family background and institutional features: the parameter of interest is the coefficient of this interaction, capturing how family background differentials vary with educational institutions. Hence, although apparently substantially different, what two-step and pooled individual models do in essence is to compare family-background

regression coefficients across educational systems.

However, models based on a single learning assessment are 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, Hanushek and Woessman (2006) propose *difference-in-difference* modeling by exploiting surveys held at different stages of the schooling career, in order to study how inequality *evolves* in early tracking countries relative to late tracking countries. This strategy allows controlling for unobserved system-level factors affecting learning inequalities already existing before the first survey. More specifically, Hanushek and Woessman (2006) use PIRLS (4^{th} grade) + PISA (age 15) to investigate the effects of tracking on reading literacy and TIMSS (4^{th} grade) + TIMSS (8^{th} grade) to investigate the effects on math. The rationale is that while in 4^{th} grade children are still in comprehensive school everywhere, in 8^{th} grade (or at age 15) they have already been tracked in some countries while in others they have not. The focus is on the effect of early tracking on the overall test scores' variability across individuals (measured by the standard deviation and selected inter-percentile ranges). Using two-step estimation, they find that in tracked systems variability increases over time relative to untracked ones, concluding that early tracking increases learning inequalities.

Drawing on this idea, a number of scholars (Waldinger 2007, Jakubowski 2010, Ammermueller, 2013; van de Werfhorst 2013) employ difference-in-difference strategies to analyze the effect of early tracking or other educational institutions on achievement differential across social origin. Interestingly, these papers reach conflicting conclusions. Similarly, Ruhose and Schwerdt (2016) use difference-in-difference to study the effect of early tracking on achievement inequalities related to migrant background. Differently from Hanushek and Woessman (2006), these scholars do not rely on two-step estimation; instead, they employ an extended version of the individual-level model, estimated on pooled data from all countries and the two assessments. The dependent variable is the test-score; explanatory variables include family background, institutional characteristics (most often, an indicator of early tracking), timing of the assessment and all two- and three-level interaction terms between these variables. The coefficient of the three-level interaction is intended to capture the extent to which family background inequalities vary over time in educational systems with certain characteristics (e.g. early tracking) relative to educational systems with other characteristics (e.g. late tracking). We will show that due to the different scaling of test scores in the different assessments, this strategy may deliver strongly biased results.

2.1 Conceptions of learning inequalities

Before moving to a detailed examination of the difference-in-difference models in the existing literature, it is useful to review how inequality is conceived and operationalized in this literature:

Overall achievement inequality

The focus is on differences among individuals, regardless of their characteristics. It can be measured by any variability index, for example the test scores' standard deviation or differences between selected percentiles of the achievement distribution (Hanushek and Woessmann, 2006).

Inequality of opportunity between family backgrounds

The focus is on average differences between children of different family backgrounds – usually conceived as social background or, less frequently, as ethnic or migratory background. It can be measured by the family background regression coefficient in a regression model with other exogenous individual characteristics as controls.

How do these two measures relate? Let γ be the family background coefficient at a given survey. In the simplest model with only one explanatory variable, under the usual OLS assumptions: $\sigma_y^2 = \gamma^2\sigma_x^2 + \sigma_\varepsilon^2$. Hence, overall inequality depends on the family-background-specific effect (γ), on the variability of family background in the population (σ_x^2), and on the influence of other factors independent of family background (σ_ε^2). This simple expression shows that overall achievement inequality and family background inequalities are distinct phenomena: indeed, they are related, but their relation need not to be strong (for related empirical evidence, see the online Supplementary Materials).

2.2 Hanushek and Woessmann's seminal paper: overall inequalities

In their seminal paper, Hanushek and Woessmann (2006) analyze the effect of early tracking on *overall achievement inequalities*, as measured by variability indexes like the scores' standard deviation. More specifically, they use two-step estimation: (i) in step-1, they estimate the *SD* in each country and at each assessment; (ii) in step-2, they examine the relation between the *SD* at *t*=2 and the institutional variable *I* given the *SD* at *t*=1. In particular, they estimate the simple linear model:

$$SD_{2c} = a + bSD_{1c} + dI_c + u_c \qquad (1)$$

where subscript *c* denotes the country and 1 and 2 index the time of the survey. *I* is the binary variable indexing early tracking and u captures country-level unobserved characteristics affecting how inequalities develop between late primary school (*t*=1) and secondary school (*t*=2).

The effect of tracking is represented by d, the average difference in the level of inequality at *t*=2 between tracked and untracked systems, given the level of inequalities already existing at *t*=1. The advantage relative to models based on single surveys is that due to conditioning on *SD* at *t*=1, unobserved factors influencing inequalities developed up to *t*=1 are taken under control. However, (1) does not control for unobserved system-level factors affecting the development of inequalities

between the two surveys. The identifying assumption is that u is orthogonal to the tracking regime; in other words, inequality changes between *t*=1 and *t*=2 should only depend on tracking or on other system-level features not correlated to the tracking regime.

Hanushek and Woessmann describe their analysis as a difference-in-difference strategy. Note however that the standard difference-in-difference refers to the double difference between the expected outcome – here, a measure of inequality – in the treated and control groups, after and before treatment:

$$DID = [E(ineq_2|I = 1) - E(ineq_2|I = 0)] - [E(ineq_1|I = 1) - E(ineq_1|I = 0)] \quad (2)$$

whereas Hanushek and Woessmann focus on:

$$DID^* = E(ineq_2|ineq_1, I = 1) - E(ineq_2|ineq_1, I = 0) \quad (3)$$

where their measure of interest is the national test scores' *SD*. Under (1), definition (3) nests the standard DID in (2) as a special case: if $b = 1$ the two definitions coincide and $DID = DID^* = d$.

2.3 Family background inequalities: pooled individual models

In the existing literature, the analyses of institutional effects on *family background* achievement inequalities follow a different modeling strategy. Individual data on different countries and assessments are pooled together, and test scores are assumed to vary with individual variables including family background, the assessment, and institutional characteristics. The strength of the family background coefficient is allowed to vary according to these institutional features. [Note that this strategy cannot be employed when inequality is conceived as a variability index, because family background differentials are expressed as differences between average performances across individuals, whereas variability indexes are not].

The simplest model is the one adopted by Waldinger (2007), Jakubowski (2010), Van de Werfhost (2013) and Ruhose, Schwerdt (2016):

MODEL M1

$$Y_{itc} = \alpha_{0c} + \alpha_1 t + \alpha_2 I_c t + \varphi X_{itc} + \xi_1 F_{itc} + \lambda_1 F_{itc} I_c + \xi_2 F_{itc} t + \lambda_2 F_{itc} t I_c + \varepsilon_{itc} \quad (4)$$

where Y is the measure of achievement, F is family background, I is the country-level binary variable indexing the early tracking regime, X is a vector of individual controls, t is a binary variable indexing the secondary school survey. Subscripts i, c and t refer to the individual, country and survey; thus, Y_{i1c} is the test score in primary school and Y_{i2c} is the test score in secondary school. Several individual (or school-level) controls and a country-level error component may also be included, but are not shown here for simplicity. The intercept is a country-specific fixed effect.

Let us denote the family background coefficients at *t*=1 and *t*=2 as γ_1 and γ_2. According to (4), $\gamma_1 = \xi_1$ in untracked and $\gamma_1 = (\xi_1 + \lambda_1)$ in tracked countries, whereas $\gamma_2 = (\xi_1 + \xi_2)$ in untracked and $\gamma_2 = (\xi_1 + \lambda_1 + \xi_2 + \lambda_2)$ in tracked countries. The identifying assumption is that the achievement gap among family backgrounds at *both* surveys may vary across countries *only* depending on the tracking regime. Instead, unobserved country-level characteristics may influence mean achievement, but may not affect family-background differentials.

Additional restrictions involving also the following model M2 are that the individual error term has the same variance across countries and that the coefficients of all other control variables are fixed across surveys and countries. This may be a substantial limitation: as shown by Guiso *et al.* (2008) and Penner (2008), for example, gender inequalities greatly differ across countries. [Limitations of pooled data models when the effects of individual variables vary across countries in standard cross-sectional analyses are discussed in Heisig *et al.* 2017].

A more flexible specification is adopted by Ammermueller (2013) to analyze the effect of tracking and other educational institutions (share of the private sector, autonomy, instruction time) on family background inequalities. The identification of institutional effects rests on the existence of variability among countries and between assessments of the institutional variables of interest. Sticking to the case of early tracking, the model can be expressed as:

MODEL M2

$$Y_{itc} = \alpha_{0tc} + \varphi X_{itc} + \xi_{1c} F_{itc} + \xi_2 F_{itc} t + \lambda_2 F_{itc} t I_c + \varepsilon_{itc} \quad (5)$$

Here the intercept may vary freely across countries and over time, and is estimated as a fixed effect by including country-time dummy variables. Family background coefficients in primary school are also unconstrained and estimated as fixed effects (hence $\gamma_{1c} = \xi_{1c}$). Instead, their variation between *t*=1 and *t*=2 depends only on institutional changes. Coefficients at *t*=2 are $\gamma_{2c} = (\xi_{1c} + \xi_2)$ for untracked and $\gamma_{2c} = (\xi_{1c} + \xi_2 + \lambda_2)$ for tracked countries. The underlying assumptions are weaker in M2 than in 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 may vary across countries only with the tracking regime *I*.

For both models M1 and M2 the parameter of main interest is λ_2, corresponding to the general standard difference-in-difference definition in (2), specified in this case as:

$$DID = \Big(E(\gamma_2|I=1) - E(\gamma_2|I=0)\Big) - \Big(E(\gamma_1|I=1) - E(\gamma_1|I=0)\Big) \quad (6)$$

representing the double difference in the family background regression coefficients between the two surveys, and between tracked (*I*=1) and untracked (*I*=0) educational systems.

Under M2 the relation between γ_2 and γ_1 can be expressed as:

$$\gamma_{2c} = \gamma_{1c} + \xi_2 + \lambda_2 I_c \quad (7)$$

so in this case λ_2 can also be interpreted as $E(\gamma_2|\gamma_1, I = 1) - E(\gamma_2|\gamma_1, I = 0)$, as in (3).

2.4 Family background inequalities: a more flexible two-level model

Models M1 and M2 appear as over-restrictive. One might consider a more flexible and transparent two-level model – let us call it M3 – with an individual-level model specified for each country and assessment and a country-level model relating regression coefficients and institutional characteristics. The coefficients of the *individual level model* of test scores Y are allowed to vary freely across countries and across assessments held at different stages of schooling:

$$Y_{itc} = \alpha_{0tc} + \gamma_{tc} F_{itc} + \varphi_{tc} X_{itc} + \varepsilon_{itc} \quad (8)$$

The regression coefficients of family background at the two assessments may depend on institutional characteristics and are related by a simple *country-level* linear model:

$$\gamma_{2c} = a + b\gamma_{1c} + dI_c + u_c \quad (9a)$$

where u captures country-level unobserved factors affecting inequalities developing between t=1 and t=2, assumed to be uncorrelated to the tracking regime represented as before by a binary indicator I. In order to allow institutional effects to vary with previous inequalities, the model could also include an interaction term:

$$\gamma_{2c} = a + b\gamma_{1c} + dI_c + g\gamma_{1c} I_c + u_c \quad (9b)$$

The effect of tracking is $d + g\gamma_1$ (reducing to d in the case of no interaction), the average difference in the family-background coefficients at t=2 between tracked and untracked systems *given* the corresponding coefficient at t=1. This definition is consistent with DID^* in (3):

$$DID^* = E(\gamma_2|\gamma_1, I = 1) - E(\gamma_2|\gamma_1, I = 0) \quad (10)$$

The identifying assumption is that inequality changes between t=1 and t=2 only depend on the tracking regime or on other system-level features not correlated to the tracking regime. Clearly, the salience of this approach depends on the existence of sufficient cross-country variability in γ_{1c} and a substantial overlap of the γ_{1c} between the subgroups of countries identified by I=0 and I=1.

2.4.1 Two-step estimation

Estimation of model M3 can be carried out by two-step estimation.

Step 1. In the first step, the family-background regression coefficients in (8) are estimated with individual level models separately for each country and assessment, so no a priori restrictions are

imposed on these coefficients over time or across countries. Since country samples are large, first step estimation usually delivers highly reliable estimates. As this specification also allows the coefficients of the control variables to vary across countries, the *F* coefficients are more likely to be valid estimates of the true family-background net effect than in pooled models M1-M2.

Step 2. In the second step, the relation between family background regression coefficients and institutions is estimated with a simple linear model at the country-level, as in (9a) or (9b). Notice that in principle second-step models can take any functional form and include other country-level explanatory variables as controls. Yet, due to small sample size, simple models with few parameters should be employed in practice. Another condition for the delivery of reliable estimates of (9a)-(9b) is the existence of sufficient variability in the γ_{1c} within institutional regimes.

2.5 Comparison of models M1-M3

Altogether, we may regard M1-M3 as all belonging to a family of models where M1 is the most restrictive and M3 is the most flexible one. In particular by comparing (7) with (9b) we see that M2 is a special case of M3 with b=1, g=0 and $u_c \equiv 0$. Hence, the first advantage of model M3 is greater flexibility. A second advantage is its transparency: first- and second- step models are simple, their underlying assumptions are clear and the interpretation of the results is straightforward. Notice that a major criticism sometimes attributed to the two-step strategy is that second step estimation is usually performed on small samples. However, although less explicit, this problem also holds for individual-data pooled models, as the relevant sample size to the estimation of regression coefficients of country-level explanatory variables is the number of countries (Wooldridge, 2010; Bryan and Jenkins, 2016).

Yet, flexibility and transparency are not the only nor the main benefits of two-step estimation over individual pooled models. In the next sections we will address an issue that has been neglected in the literature adopting difference-in-difference strategies with international assessments: we will show that when test scores are measured on different scales over time – as occurs in international assessments held at different stage of schooling – two-step estimation of M3 delivers meaningful results whereas pooled individual models M1 and M2 in general *do not*.

2.6 Extension to other educational institutions

Early tracking is by far the institution that has received the greatest attention in terms of its potential impact on achievement inequalities. The role played by other features of the education systems has been analyzed in a similar perspective, either with single international assessments, or by exploiting two assessments with difference-in-difference strategies. As mentioned above, Ammermueller (2013) uses an extended version of model M2 to analyze the effect on inequalities of the strength of the private sector, the degree of autonomy and time devoted to instruction.

Indeed, the case of tracking is ideal to be analyzed with a difference-in-difference design, because in 4th grade children are still in comprehensive school in all countries, whereas in at age 15 tracking has already been enforced in some countries but not in others, allowing studying how inequality evolves in early tracking countries relative to late tracking countries. Nonetheless, the models described in sections 2.3 and 2.4 can be adapted to study the effect of other institutional features described by non-binary variables and varying over time and across countries with different patterns (for details, see Appendix A.1). The discussion on the validity of the difference-in-difference individual pooled models developed in the following sections also applies to them.

3. Comparing learning inequalities as children grow: the scaling issue

The core question when evaluating the effect of early tracking on family background inequalities with difference-in-difference strategies is: *Do family background differentials in achievement increase more (or decrease less) in tracked systems relative to untracked systems?* Hence, we face the problem of assessing how inequalities develop as children grow older in different systems.

We start by saying that we will not address issues related to the tests' constructs. Scholars usually utilize TIMSS math test scores in 4th and 8th grade, designed by IEA (the *International Association for the Evaluation of Educational Achievement*) to measure curricular competencies, or PIRLS and PISA's reading test scores that, despite being administered by different agencies (IEA and OECD), are considered to follow similar constructs (Zuckerman *et al.*, 2013). Instead, we focus on the fact that test scores in international assessments are not 'vertically equated', i.e. achievement is not measured on the same scale at different grades. As discussed by Bond and Lang (2013), scaling issues in test scores make it difficult to analyze the development of average test score differentials over time.

Our line of reasoning can be summarized as follows: If expressed in different metrics, cross-sectional regression coefficients are not comparable across surveys: the difference $(\gamma_2 - \gamma_1)$ is meaningless. Although this is a rather trivial point, we show it formally under a stylized structural achievement growth model that will set the basis for an in-depth examination (carried out in Section 4) of the results delivered by the difference-in-difference strategies employed in the literature. For some reason, the scaling issue has been ignored in this literature: we presume that the implicit assumption is that the problem would disappear when applied to the double difference $(\gamma_2 - \gamma_1|I = 1) - (\gamma_2 - \gamma_1|I = 0)$. In Section 4 we will prove that this is generally not the case.

3.1 Test scores in international assessments

Let us start with a brief introduction on how test scores in international assessments are produced. International surveys rely on Item Response Theory (IRT). These methods take into account the items' difficulty, and in some cases the guessing probability and the items' discriminatory power. 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. 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. This is a crucial element of international assessments, because it allows comparing the educational outcomes across countries. Transformed scores have mean equal to 500 and SD equal to 100 (OECD 2009). Let us call these scores *original scores* (standardized across countries). We may also consider *within-countries* standardized scores, produced by standardizing original scores relative to each country's mean and SD. Patently, the achievement of individuals from different countries is comparable only with original scores.

Inequalities within countries

Focusing on inequalities *within countries*, if we compare two individuals from country A or two individuals from country B with original PISA scores, we observe how many SD they are apart with respect to the cross-country SD in PISA. Using within-countries standardized scores, if we compare two individuals from country A we observe how many SD they are apart with respect to the SD of country A; if we compare two individuals from Country B we observe how many SD they are apart with respect to the SD of country B. To fix idea, consider the following example, relative to PISA (fictional data):

Table 1. Original scores and within countries standardized scores

	Original scores					(Within-countries) standardized scores				
	Mean	*SD*	*F*=1	*F*=0	*F*-difference	Mean	*SD*	*F*=1	*F*=0	*F*-difference
Country A	500	100	560	440	120	0	1	0.6	-0.6	1.2
Country B	500	70	542	458	84	0	1	0.6	-0.6	1.2

Let F represent family background (F=1 is high background and F=0 is low background). Using original scores, the F-difference tells us that family background inequalities are larger in country A than in country B, because the difference is 120 points in the former and only 84 in the latter. The F-difference using within-countries standardized scores tells us that family background inequalities have the same relative weight (relative to overall inequality) in countries A and B. Clearly, these results do not convey the same information. [To our knowledge, all papers in the literature analyzing inequalities use original scores.]

Inequalities at different stages of schooling

If we wish to analyze the evolution of inequalities at different stages of schooling, we have to consider comparing test scores' measures of inequality *across assessments.* A relevant distinction in this case is between *vertically equated* and *non-equated* tests. In equated tests, some items appear in both assessments, allowing their "anchoring" (Bond and Lang, 2013). This enables to express test scores in a common metric and evaluate achievement growth. However, international assessments held at different grades/age are not equated. As a result, as we discuss below, comparing achievement inequalities over time is generally not meaningful with original scores, and conveys only limited information on the evolution of inequalities when using (within countries) standardized scores.

3.2 The scaling issue

A simple structural achievement growth model

Consider a simple 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 (Contini and Grand, 2017). This can be viewed as an ideal model of cognitive ability, assuming that ability can be measured on a meaningful interval scale and that it evolves linearly. Initial ability and growth may be affected by ascribed individual characteristics such as family background (e.g. socioeconomic status, minority, ethnic or immigrant origin) or gender.

Suppose we have two cross sectional surveys assessing students' learning in a given country 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 test scores are measured on the *same* scale in the two assessments. Let y_{i2} be the score of individual i at t=2 and y_{i1} her score at t=1. To simplify the exposition, we refer to a single explanatory variable F (but clearly other individual controls should be included) and assume that:

$$y_{i1} = \mu_1 + \rho F_i + \varepsilon_{i1} \tag{11}$$

In our current example, F is an indicator of family background, with F=1 for high background and F=0 for low family background. Achievement at t=2 is given by achievement at t=1 plus achievement growth δ:

$$y_{i2} = y_{i1} + \delta_i \tag{12}$$

Growth may be assumed to depend linearly on explanatory variables and may also depend on previous achievement:

$$\delta_i = \Delta + \beta F_i + \theta y_{i1} + \varepsilon_{i2} \tag{13}$$

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 (new inequalities developed between the two assessments). Instead, θ captures carry-over effects of pre-existing inequalities.

When scales are different

With longitudinal data and achievement measured on the same scale at different grades, it is possible to evaluate achievement growth for each child, estimate model (13), and identify the structural parameters β and θ (thus, disentangle the two different mechanisms responsible of how inequalities develop over time).

However, international learning assessments are cross-sectional (this occurs also for many national assessments). Moreover, achievement is measured on different scales as children grow older. In this case, we have to distinguish between the (unknown) scores y_1 representing achievement at *t*=1 according to the scale employed at *t*=2, and observed scores y'_1. Assuming for simplicity a linear relation between these scales (where φ and ω are unknown and unidentifiable):

$$y_{i1} = \varphi + \omega y'_{i1} \tag{14}$$

from (11) we obtain the model relating observed scores y'_{i1} to family background F:

$$y'_{i1} = const + \frac{\rho}{\omega} F_i + \frac{\varepsilon_{i1}}{\omega} \tag{15}$$

so the estimable F-regression coefficient at *t*=1 is $\frac{\rho}{\omega}$ and represents the total family background differential developed up to *t*=1 in the metric of the first assessment.

Substituting (13) and (14) into (12), we obtain the dynamic model for y_2 as a function of y'_1:

$$y_{i2} = y_{i1} + \delta_i = (\varphi + \omega y'_{i1}) + \Delta + \beta F_i + \theta(\varphi + \omega y'_{i1}) + \varepsilon_{i2} \tag{16}$$

Clearly, this model cannot be estimated with cross-sectional data. Further substituting (15) into (16), we obtain the *cross-sectional model* for y_2:

$$y_{i2} = const + (\beta + (1+\theta)\rho) F_i + (1+\theta)\varepsilon_{i1} + \varepsilon_{i2} \tag{17}$$

where the F-regression coefficient is $(\beta + (1+\theta)\rho)$ and represents the total family background achievement differential developed up to *t*=2, in the metric of the second assessment.

In conclusion, the difference between the estimable cross-sectional regression coefficients at *t*=2 and *t*=1 is:

$$(\beta + (1+\theta)\rho) - \frac{\rho}{\omega} \tag{18}$$

If test scores are measured on different scales ($\omega \neq 1$) and ω is unknown and not identifiable – as occurs if tests are not equated – this quantity delivers meaningless results.

Standardized test 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 regression coefficients of models run on standardized scores. These differentials are invariant to the score metric, hence are comparable:

$$E(z_1|F=1) - E(z_1|F=0) = \frac{\rho/\omega}{\sigma_{y_1'}} = \frac{\rho}{\sigma_{y_1}} \quad (19)$$

$$E(z_2|F=1) - E(z_2|F=0) = \frac{(1+\theta)\rho+\beta}{\sigma_{y_2}} \quad (20)$$

The difference between (20) and (19) informs on how many standard deviations two individuals of different family backgrounds are apart at *t*=2 as compared to *t*=1. However, standard deviations at *t*=1 and *t*=2 are generally different, so the sources of the observed change are unclear. In fact, children's achievement is not influenced only by family background: if the test-scores' variability increases because of growing differentials related to other characteristics (e.g. increasing gender inequalities), we could observe *decreasing* family background inequalities even if $\theta > 0$ and $\beta > 0$. [However, it can be shown however that a positive difference between (20) and (19) implies β>0].

Hence, even if regression coefficients on standardized scores are comparable, their difference is not fully informative on the evolution of family background inequalities as children grow.

4. The scaling issue in difference-in-difference strategies

Let us summarize the main points raised so far:

(i) In Section 2 we reviewed the difference-in-difference strategies employed in the literature and highlighted that, in essence, individual pooled models identify the effect of institutions on family background inequalities by taking the (double) difference of cross-sectional regression coefficients relative to assessments administered at different children's age.

(ii) In Section 3 we specified a structural achievement growth model and highlighted that, even under this stylized model, when test scores are non-equated, the difference of cross-sectional regression coefficients based on original scores is generally meaningless, and the difference based on standardized scores conveys limited information on the evolution of family background inequalities as children age.

Against this background, we now draw implications on the validity of the difference-in-difference strategies employed in the literature using international assessments delivered at different stages of schooling (presented in sections 2.3 and 2.4), by expressing *DID* as a function of the structural parameters of the achievement growth model. Notice that although we stick to the example of early tracking, the argumentation can be easily extended to other institutional variables not represented by a binary variable and that may vary over time and across countries with different patterns.

To fix ideas, think of PIRLS (4th grade) as the assessment at t=1 and PISA (age 15) as the assessment at t=2. Data are cross-sectional and test scores are not equated. Following the structural model, achievement depends on family background at t=1 and t=2 according to (15) and (17). Thus, regression coefficients, in the most general setting variable across countries, may be expressed as:

$$\gamma_{1c} = \rho_c/\omega \qquad \text{at } t=1 \qquad (21)$$
$$\gamma_{2c} = \beta_c + (1+\theta_c)\rho_c \qquad \text{at } t=2$$

where ω reflects the different scale used to measure test scores in the two surveys.

4.1 Difference-in-difference with pooled individual models (M1 and M2)

As shown on Section 3, the difference between regression coefficients γ_2 and γ_1 when test scores are not equated is generally meaningless. The question now is: *Does the double differentiation of regression coefficients solve the scaling problem?* We now show that the answer is no.

Difference-in-difference with model M1

In model M1 (section 2.3), regression coefficients are allowed to vary across countries only according to institutional features. Restricting all structural parameters accordingly, and substituting (21) into the standard DID: $\Big(E(\gamma_2|I=1) - E(\gamma_2|I=0)\Big) - \Big(E(\gamma_1|I=1) - E(\gamma_1|I=0)\Big)$ we estimate:

$$DID = [(\beta_T + (1+\theta_T)\rho_T) - (\beta_U + (1+\theta_U)\rho_U)] - [(\rho_T - \rho_U)/\omega] \qquad (22)$$

The first term in square brackets is the difference between the regression coefficients in tracked (*T*) and untracked systems (*U*) in secondary school; the second one is the difference between the regression coefficients in tracked and untracked systems in primary school. This expression (corresponding to coefficient λ_2) delivers meaningful results only in very peculiar circumstances: (i) in the fortuitous case that the different scales employed to measure achievement in the two assessments were additively related ($\omega = 1$); (ii) in the fortuitous case that the degree of inequality at t=1 happened to be equal in tracked and untracked systems ($\rho_T = \rho_U$); (iii) in the fortuitous case that the degree of inequality at t=2 happened to be equal in tracked and untracked systems, i.e. if $\beta_T + (1+\theta_T)\rho_T = \beta_U + (1+\theta_U)\rho_U$. In general, however, the effect of tracking ends up being

estimated by the difference between non-comparable quantities: the double differentiation does *not* solve the scaling problem (see also the simulation exercise in Appendix A.1).

Difference-in-difference with model M2

In model M2 (section 2.3) inequalities at *t*=1 are unconstrained, whereas the changes occurring between *t*=1 and *t*=2 may only depend on the tracking regime. For this reason we let ρ vary freely across countries (indicated as ρ_c), but constrain β and θ to depend on tracking. Substituting the corresponding regression coefficients into the expression for the standard DID we obtain:

$$DID = [(\beta_T + (1+\theta_T)E_T(\rho_c)) - (\beta_U + (1+\theta_U)E_U(\rho_c))] - [E_T(\rho_c) - E_U(\rho_c)]/\omega \quad (23)$$

where $E.(\rho_c)$ is the expected value of ρ in a given tracking regime. Once again, the estimated DID depends on the unknown scaling factor ω and delivers meaningful results only under the fortuitous circumstances described above for M1 (see also the simulation exercise in Appendix A.1).

4.2 Difference-in-difference with two-step estimation (model M3)

Since model M3 (section 2.4) is more general than M1 and M2, the conclusion that when test scores are not equated the estimated standard *DID* delivers generally meaningless results applies to M3 as well. So, is there an advantage in using M3? The crucial point is that in model M3 the identification of the institutional effects is not reached by estimating the standard DID, but $DID^* = E(\gamma_2|\gamma_1, I=1) - E(\gamma_2|\gamma_1, I=0)$. This is accomplished by estimating second step models (9a)-(9b), that directly relate regression coefficients at *t*=2 with the tracking regime and regression coefficients at *t*=1. The scaling problem *disappears* because dependent and independent variables in a regression model need not to be on the same scale.

Against this background, we may wish to analyze how two-step estimates relate to the structural parameters of the achievement growth model. According to (21), the following holds:

$$E(\gamma_{2c}|\gamma_{1c}) = E(\beta_c) + \omega\big(1 + E(\theta_c)\big)\gamma_{1c} \quad (24)$$

Let us recall second-step models (9a) and (9b) in M3:

$$\gamma_{2c} = a + b\gamma_{1c} + dI_c + u_c$$

$$\gamma_{2c} = a + b\gamma_{1c} + dI_c + g\gamma_{1c}I_c + u_c$$

Equation (24) is consistent with the first specification if on average β (new family background inequalities developed between *t*=1 and *t*=2) varies across countries with the tracking regime and θ (carry-over effect of previously established inequalities) does not vary with the tracking regime. It is consistent with the second if both β and θ vary with the tracking regime.

Thus, in principle second-step estimation allows to draw conclusions on the mechanisms underlying how family background inequalities change over time. More specifically: a resulting $d \neq 0$ suggests that β varies between tracked and untracked regimes. Instead, $g \neq 0$ suggests that θ varies between tracked and untracked regimes. In fact, even if $E(\theta)$ is not identified when ω is unknown (i.e. when tests are not equated), $\omega E(\theta)$ is identified and the expression $\omega\big(1 + E(\theta_c|I = 1)\big) \gtrless \omega\big(1 + E(\theta_c|I = 0)\big)$ implies $E(\theta_c|I = 1) \gtrless E(\theta_c|I = 0)$.

Still, caution is advised when interpreting two-step results in this manner, as the intercept's estimate is usually unstable with small samples and the linear specification may be only a convenience approximation of a potentially more complex relation between previous and later achievement gaps.

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 (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.

Following Abadie *et al.* (2015) who argue that a careful choice of the countries is necessary to reduce the risk of unobserved country level confounding factors, we consider only European and Anglo-Saxon countries, as they share comparable schooling systems, societal organization and cultures, ending up with 24 countries participating to both assessments.

By tracking, we refer to the formal sorting process into schooling institutions providing 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 up to age 15, as "untracked" otherwise. In our sample, we have 12 tracked and 12 untracked countries (Table 1). However, we also carry out robustness checks with alternative tracking variables: a dummy classifying countries tracking at age 15 as untracked (since tracking has taken place very recently) and the number of years since tracking.

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 2. Countries in the empirical analysis by tracking regime

COUNTRIES	AGE OF TRACKING	DUMMY TRACKING	COUNTRIES	AGE OF TRACKING	DUMMY TRACKING
Austria	10	1	Canada	18	0
Belgium	14	1	Denmark	16	0
Bulgaria	14	1	Latvia	16	0
France	15	1	Lithuania	16	0
Germany	10	1	New Zealand	16	0
Hungary	10	1	Norway	16	0
Israel	15	1	Poland	16	0
Italy	14	1	Romania	16	0
Luxembourg	12	1	Russian Fed.	16	0
Netherlands	12	1	Spain	16	0
Slovakia	11	1	Sweden	16	0
Slovenia	15	1	USA	18	0

NOTE. Source: see Appendix B, Table B.2.
Dummy tracking: =1 if tracking occurs at age<=15, 0 otherwise
Alternative definition (see Robustness checks in Appendix C): Dummy tracking: =1 if tracking at age<=14, 0 otherwise

In line with the methodological considerations developed in the previous sections, we apply two-step analysis to overall inequalities (replicating the analyses carried out in Hanushek and Woessmann's (2006) with test scores' standard deviations on more recent data and a different set of countries) and to social background inequalities. In the first step, for each country and assessment we estimate the test scores standard deviations and the social background regression coefficients with model (8). 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. We focus on the linear combination of these coefficients (under the heading F-GAP), highlighting the effect of tracking on the test-scores differential 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 (see Appendix B for the definition of individual-level variables). In the second step, we analyze the relationship between estimated inequalities at t=2 and the tracking regime, given inequalities at t=1.

5.2 First step results: preliminary findings

First-step regressions are run with R routines designed to handle plausible values and complex sampling (Caro and Biecek 2017), using student replicate weights. The full set of first step results is available in the online Supplementary Materials.

Focusing on overall inequality, we find that on average the SD at *t*=1 (PIRLS) is slightly larger in untracked than in tracked countries, whereas the relation reverts at *t*=2 (PISA), where tracked countries display larger values (Table 3). A similar pattern holds when looking at social background inequalities, as the average achievement gap between high and low strata (F-GAP) is nearly the same at *t*=1, while at *t*=2 it becomes much larger in tracked countries. Acknowledging that the interval scale of test scores is sometimes questioned (Bond and Lang 2013), we also look at country rankings – from smallest lo largest – obtaining similar results, but even more marked.

Table 3. Country-level absolute measures of inequality and rankings

	Original scores				Country rankings			
	SD1	SD2	F-GAP1	F-GAP 2	SD1	SD2	F-GAP 1	F-GAP2
Tracked	67.2 (12.4)	97.3 (9.5)	83.5 (19.2)	134.5 (22.3)	11.3 (7.9)	15.5 (6.5)	13.0 (7.3)	16.3 (5.9)
Untracked	69.5 (9.4)	90.3 (6.1)	83.4 (22.4)	114.0 (13.5)	13.7 (6.3)	9.5 (6.7)	12.0 (7.1)	8.7 (6.1)

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

Under the heading F-GAP1 and F-GAP2 we report results relative to the effect of tracking on the difference between tertiary educated parents with the largest number of books and no tertiary educated parents with the lowest number of books: F-GAP=[ln(500)*β(*ln*(n° books))+β(tertiary degree)]- ln(5)*β(*ln*(n° books).

5.3 Second step estimation

To analyze *overall inequalities*, we replicate Hanushek and Woessman's analyses and estimate model (1), as well as an extended version of this model including an interaction term between previous inequalities and the tracking regime. To analyze *social background inequalities*, we estimate models (9a) and (9b) relating the country-level measures of family-background inequality at *t*=2 to the tracking regime, given inequality at *t*=1. Results are summarized in Table 4. The coefficients of prior inequalities are always positive, indicating that countries with high inequalities in primary school also tend to have high inequalities in secondary school.

Findings on overall inequalities – columns (1) and (2) show that given SD in primary school, the SD is higher on average in tracking countries. The interaction effect is positive (meaning that the SD tends to increase more between primary and secondary school in tracking countries relative to untracked countries) but not statistically significant. On average, the SD at *t*=2 is 8 point higher (i.e. 8% of the average national SD) in tracked countries relative to untracked countries with the same SD at *t*=1. Our results are consistent with the results in Hanushek and Woessman (2006), although they report substantially larger effects of early tracking (almost a quarter of a SD for reading literacy).

Findings on the effects of tracking on social background inequalities – columns (3) and (4) – indicate that early tracking is associated to an increase in inequalities related to social background. Given educational inequality already existing in primary school, the linear combination of the two

social background regression coefficients at age 15 (the F-GAP) is on average 20.4 score units – 0.204 standard deviations in the OECD distribution – higher in tracked than in untracked systems. Adding the interaction term shows that the difference between tracked and untracked countries tends to increase at higher levels of inequality at *t*=1. Similar results are found when considering countries tracking at age 15 as untracked (see robustness checks in the Appendix C, Table C.1), whereas no interaction effect is observed when considering the number of years since tracking (Appendix C, Table C.2).

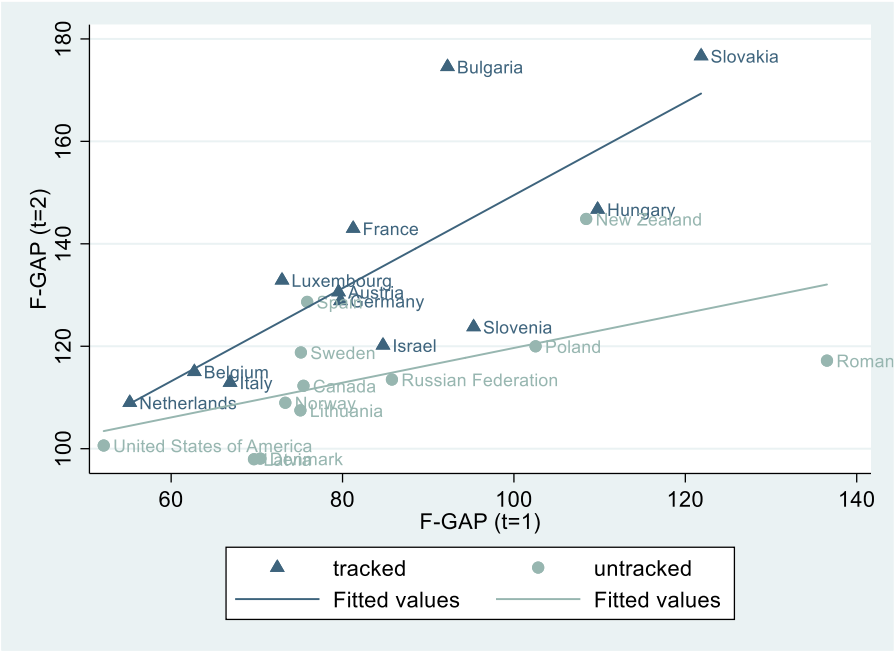
Table 4. Second step results. Cross-country regression models

	INEQUALITY MEASURE			
	Overall inequalities		Social background inequalities	
	SD		**F-GAP**	
	(1)	(2)	(3)	(4)
Constant	60.43***	78.55***	65.72***	85.77***
Tracking regime	8.01***	-20.17	20.40***	-27.09
Inequality measure at *t*=1	0.430***	0.170	0.579***	0.339*
Tracking* Inequality measure at *t*=1		0.410		0.569*
N countries	24	24	24	24
R^2	*0.468*	*0.530*	*0.573*	*0.649*

NOTES. Under the heading SD we report results relative to the effects of tracking on the country standard deviation. Under the heading F-GAP we report results relative to the effects of tracking on the difference between tertiary educated parents with the largest number of books and no tertiary educated parents with the lowest number of books: F-GAP=[ln(500)*β(*ln*(n° books))+β(tertiary degree)]- ln(5)*β(*ln*(n° books).
Columns (1) and (3) refer to the model with no interaction; columns (2) and (4) to models with interaction between the tracking indicator and inequality at *t*=1.
* 0.05<p-value<0.10, ** 0.01<p-value<0.05, *** p-value<0.01

In Figure 1 we show the scatter diagram depicting observed social background inequalities (the F-GAP) at *t*=2 against the corresponding values at *t*=1. The straight line represents the predicted relation by tracking regime, according to the estimates of model (9b) reported in column (4) of Table 4. First of all, this graph shows that in primary school social background inequalities vary considerably across countries even within tracking regimes. Secondly, it shows that at low levels of social background inequality in primary school, there is little difference in secondary school inequalities between countries with and without tracking; instead at high levels of inequality in primary school, tracked systems tend to become more unequal than untracked systems.

Figure 1. Social background effects at *t*=2 and *t*=1



NOTE. F-coefficients in primary (PIRLS) and secondary school (PISA)- estimates of model (9b).

We now attempt to interpret the results relative to the effect of tracking on social background inequalities in terms of the structural parameters of the achievement growth model (as described in Section 4.2). Since the intercept does not significantly differ from 0, we should conclude that 'new inequalities' developed between primary and secondary schools given prior achievement, represented by β, are similar in tracked and untracked systems (or perhaps even lower in tracked systems, since the point estimate is negative). Instead, carry-over effects of previous social background inequalities in achievement, represented by θ, seem to be substantially larger in tracked countries than in untracked countries, as implied by the substantially higher slope estimated for the former. In other words, according to this interpretation, the reason why social background inequalities tend to widen between primary and secondary school in tracked systems relative to untracked systems, is because the gap between well- and poor- performing children in primary school (already socially determined) widens more in the former as compared to the latter. This seems reasonable: in tracked systems, well-performing children tend to attend the academic track and low-performing children tend to choose more labor-market oriented schools, although with different probabilities across social backgrounds. Thus, the gap between well- and poor- performing children may widen more sharply in these countries than in comprehensive school systems. As already remarked, however, due to the small sample size in step 2 estimation, this structural interpretation of the results is to be taken with caution.

5.4 Difference-in-difference with pooled individual regression models

For illustrative purposes, we also 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 gender and age as controls. The model was run on a total of 240,273 individuals taking either the PIRLS or the PISA test, in the 24 countries of Table 2. The *DID* estimate turns out to be 22.50 (significant at the 0.10 level) for M1 and 24.83 (significant at the 0.05 level) for M2. Interestingly, these estimates are not sharply different from those delivered by the two-step estimation model (9a). The reason is that in this particular case inequalities at t=1 are very similar on average in the two regimes: as shown in Table 3 the mean F-gap is 83.5 points in tracked countries and 83.4 in untracked countries. Thus, in this case we fall into one of the *fortuitous* circumstances discussed in section 4.1 where results of M1 and M2 – although delivered by unnecessarily restrictive and untransparent models – *happen* not to be meaningless, as the estimated *DID* ends up being expressed in the metric of test scores at t=2.

6. Concluding remarks

This article aims at contributing to the literature that reflects on the correct use of international learning assessments in econometric modelling (e.g. Jerrim *et al.*, 2017). The specific purpose of this paper is to provide an in-depth analysis of difference-in-difference strategies aimed at evaluating the effect of institutional features on learning inequalities, exploiting international assessments administered to children of different age. In the existing literature, difference-in-difference has been carried out with two-step estimation by Hanushek and Woessmann (2006), who analyzed the effect of early tracking on overall inequalities (captured by test score variability indexes). Other scholars, instead, analyzed the effect of early tracking and other features of the educational system on family background inequalities (captured by the family-background regression coefficients), using individual level models on data pooled from different countries and different assessments. We demonstrate that scaling issues entailed by using non-equated test scores at different stages of schooling may severely undermine the validity of the results delivered by pooled individual level models. Scaling issues do not apply instead to two-step estimation. Hence, provided that difference-in-difference be reputed a valid strategy for the problem at stake, we view two-step estimation as a much better alternative to pooled models' estimation. Our methodological discussion can be extended to different research areas: the scaling issue may be relevant when analyzing the impact of policies on fundamental individual characteristics changing over the life course, other than learning – for example, health, well-being or life satisfaction – for which different measurement tools are needed as people grow up from early childhood to adulthood (Lippman *et al*. 2011).

In the empirical section of the paper, we analyze the effect of early tracking on inequalities in reading literacy. Consistently with the methodological discussion, we apply two-step analysis on both overall achievement inequalities and social background inequalities. Our findings are that, given inequality in primary school, inequalities in secondary school are substantially larger in early tracking than in late tracking countries. When focusing on social background inequalities, we find that the difference between tracking regimes increases with inequality in primary school: early tracking seems to be detrimental to equity in particular in countries with strong inequalities already existing in primary school. Results on overall inequalities (measured by test scores' standard deviations) go in the same direction, but are somewhat weaker. Altogether, our evidence is that early tracking increases achievement inequalities, in particular by widening the difference between children of different social origin. Pushing our conclusions even further, there is some evidence that the reason why social background inequalities tend to widen in tracked relative to untracked systems between primary and secondary school, is not related to a larger gap developed within this time span between previously equally performing children of different social origin, but instead to different carryover effects of inequalities already existing in primary school. More research is needed to confirm these findings and provide a fully convincing interpretation for them.

A remark on the limitations of policy evaluations based on cross-country analyses is also in order. In general, results are not easily interpretable in causal terms. The main reason is that countries vary on a multitude of characteristics, so it is difficult to 'hold other things constant'. This criticism applies in particular to conventional cross-section analyses, but despite milder conditions required, it may be directed also to difference-in-difference models. Another reason is sample size, because identification of policy effects is reached by exploiting cross-country variability in institutional variables, and the number of countries is usually small. In spite of these limitations, it is only by gathering evidence from different contexts and analytical strategies that we can make general statements on the effects of the policies or institutions of interest. Since institutions/policies are rarely subject to reforms (and if they do, it is 'by luck'), we think it would be unwise not to exploit the great opportunity provided by international standardized learning assessments to build knowledge on how schooling policies and institutional arrangements relate to educational outcomes. Yet, modelling strategies have to be transparent, as well as the underlying assumptions and the conditions for the validity of the results.

References

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

Abadie A., A. Diamond, J. Heinmueller (2015) Comparative politics and the synthetic control method, *American Journal of Political Science*, 59(2), 495–510

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.

Bol T., J. Witschge, H. G. Van de Werfhorst, J. Dronkers (2014) Curricular tracking and central examinations: counterbalancing the impact of social background on student achievement in 36 countries, *Social Forces*, 92(4), 1545-1572

Bond T., K. Lang (2013) The Evolution of the Black-White Test Score Gap in Grades K–3: The Fragility of Results *The Review of Economics and Statistics*, 95(5), 1468-1479

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

Brunello G., D. Checchi (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.

Caro D. H., P. Biecek (2017). "intsvy: An R Package for Analyzing International Large-Scale Assessment Data." *Journal of Statistical Software*, **81**(7), 1–44.

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.

Chmielewski A. K., Reardon, S.F. (2016) Patterns of cross-national variation in the association between income and academic achievement, *AERA Open* 2(3): 1-27

Contini D., E. Grand (2017). On estimating achievement dynamic models from repeated cross-sections, *Sociological Methods and Research*, 46, 4, 683–714

De Gregorio J. and J-W. Lee (2002) Education and income inequality: New evidence from cross country data, Review of Income and Wealth, 48, 3, 395-416

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

Green A., J. Preston and J. Janmaat (2006) *Education, Equality and Social Cohesion. A Comparative Analysis*, Palgrave Macmillan, New York

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.
Hanushek, E.A., Woessmann L. (2015) *The Knowledge Capital of Nations: Education and the Economics of Growth*, CESifo Book Series, MIT Press

Heisig, J.P., M. Schaeffer and J. Giesecke (2017). The costs of simplicity: Why multilevel models may benefit from accounting for cross cluster differences in the effects of controls, *American Sociological Review*, 82(4), 796–827

Horn D. (2009) Age of selection counts: a cross-country analysis of educational institutions, *Educational Research and Evaluation*, 15(4), 343–366

Jackson M. eds (2013) *Determined to succeed? Performance versus choice in educational attainment*, Stanford University Press, Stanford CA

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

Jerrim, J., L. A. Lopez-Agudo, O. D. Marcenaro-Gutierrez, N. Shure (2017). What happens when econometrics and psychometrics collide? An example using the PISA data, *Economics of Education Review*, 61, 51-58

Kerr S. P., T. Pekkarinen, R. Uusitalo (2013) School tracking and development of cognitive skills, *Journal of Labor Economics*, 31(3), 577-602

Lippman H,. K. Anderson Moore, H. McIntosh (2011) Positive Indicators of Child Well-Being: A Conceptual Framework, Measures, and Methodological Issues, *Applied Research Quality Life*, 6, 425-449

Malamud O., C. Pop-Eleches (2011) School tracking and access to higher education among disadvantaged groups, *Journal of Public Economics*, 95 (11-12), 1538-1549

Meghir C., Palme M. (2005) Educational Reform, Ability, and Family Background, *American Economic Review*, 95(1) 414-424

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.

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

OECD (2009) *PISA Data Analysis Manual: SPSS and SAS*, Second Edition

OECD (2014*) PISA 2012 results in focus. What 15-year-olds know and what they can do with what they know*, http://www.oecd.org/pisa/keyfindings/pisa-2012-results-overview.pdf

Penner A.M. (2008) Gender differences in extreme mathematical achievement: an international perspective on biological and social factors, *American Journal of Sociology* 114(S1), S138-S170

Piopiunik M. (2014) The effects of early tracking on student performance: evidence from a school reform in Bavaria, *Economics of Education Review*, 42, 12-33

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 (2016) Does early educational tracking increase migrant-native achievement gaps? Difference-in-difference evidence across countries, *Economics of Education Review*, 52, 134–154.

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

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.

Woessmann L. (2016) The importance of school systems: Evidence from international differences in student achievement, *Journal of Economic Perspectives* 30(3), 3-32

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

Appendix A. Methodological materials

A.1 Difference-in-difference models applied to generic educational institutions

Individual pooled models

Ammermueller (2013) analyzes the effect of tracking and of other educational institutions (share of the private sector, autonomy, instruction time) on family background inequalities. Identification of institutional effects rests on the variability among countries and between assessments of these institutional variables *I*. In the general case, the model can be expressed as:

$$Y_{itc} = \alpha_{0tc} + \varphi X_{itc} + \xi_{1c} F_{itc} + \xi_2 F_{itc} t + \lambda_2 F_{itc} t (I_{2c} - I_{1c}) + \varepsilon_{itc}$$

The institution of interest *I* here is not necessarily described by a binary variable (consider for example the share of private sector, that can take any value from 0% to 100%), and may vary across stages of schooling in different ways.

Family background coefficients in primary school are unconstrained and estimated as fixed effects (hence $\gamma_{1c} = \xi_{1c}$). Instead, the variation between *t*=1 and *t*=2 depends only on institutional changes. Coefficients at *t*=2 are $\gamma_{2c} = (\xi_{1c} + \xi_2)$ when $I_{2c} = I_{1c}$ (i.e. when there are no changes between *t*=1 and *t*=2) and $\gamma_{2c} = \xi_{1c} + \xi_2 + \lambda_2 (I_{2c} - I_{1c})$ with institutional changes.

The underlying assumptions are that unobserved country characteristics are allowed to affect family background inequalities at *t*=1; instead, the *difference* in family background inequalities between *t*=1 and *t*=2 may be influenced only institutional changes occurring in the meantime.

The parameter of main interest is λ_2, corresponding to:

$$\big(E(\gamma_2 | I_2 - I_1 = i + 1) - E(\gamma_2 | I_2 - I_1 = i)\big) - \big(E(\gamma_1 | I_2 - I_1 = i + 1) - E(\gamma_1 | I_2 - I_1 = i)\big)$$

representing the effect of a one-unit institutional change on family background inequalities.

Two-step models

The coefficients of the *individual level model* of test scores *Y* are allowed to vary freely across countries and across assessments held at different stages of schooling:

$$Y_{itc} = \alpha_{0tc} + \gamma_{tc} F_{itc} + \varphi_{tc} X_{itc} + \varepsilon_{itc}$$

The (random) regression coefficients of family background at the two assessments depend on institutional characteristics and may be related by a simple *country-level* linear model:

$$\gamma_{2c} = a + b\gamma_{1c} + d(I_{2c} - I_{1c}) + u_c$$

where $(I_{2c} - I_{1c})$ capture institutional changes occurring between *t*=1 and *t*=2 and u captures country-level unobserved factors affecting inequalities developing in the same time span, assumed to

be uncorrelated to institutional changes. The effect of tracking is d, the average difference in the family-background coefficient at *t*=2 that can be ascribed to an additional one-unit change in the institutional variable, *given* the coefficient at *t*=1:

$$DID^{*} = E(\gamma_2|\gamma_1, I_2 - I_1 = i + 1) - E(\gamma_2|\gamma_1, I_2 - I_1 = i)$$

The identifying assumption is that inequality changes between *t*=1 and *t*=2 only depend on the institutional variable of interest or on other system-level features not correlated to it.

A.2 DID in pooled individual models: a simulation exercise

To fix ideas, we propose an example showing what is actually estimated with individual pooled models M1 and M2, when test scores are equated and when they are not equated. To illustrate the issue, we impose a rather large difference between parameters in tracked and untracked systems (yet, results are qualitatively the same when using other sets of parameters).

We simulate data for 30 countries, half of which are tracked and the other half untracked. There are two assessments. Each country has 3000 observations per assessment. We first simulate test scores according to equations (11)-(13):

$$y_{i1} = \mu_1 + \rho F_i + \varepsilon_{i1} \quad (11)$$

$$y_{i2} = y_{i1} + \delta_i \quad (12)$$

$$\delta_i = \Delta + \beta F_i + \theta y_{i1} + \varepsilon_{i2} \quad (13)$$

as if achievement was expressed in the same metrics at the two assessments, and then transform scores at *t*=1 into a different metrics as in (14):

$$y_{i1} = \varphi + \omega y'_{i1} \quad (14)$$

with scaling factor $\omega = 0.5$ and $\varphi = 0$.

We let ρ (representing inequalities at *t*=1) vary across countries according to a uniform distribution: $U(40{,}60)$ in tracked countries (average=50), $U(20{,}40)$ in untracked countries (average=30). Instead, β and θ (representing inequalities developing between *t*=1 and at *t*=2) are fixed within tracking regime. The parameters and the corresponding γ (family background regression coefficients) are described in the upper panel of Table A1. We distinguish between the latent γ_1, – computed as if test scores at *t*=1 and *t*=2 were equated, hence expressed on the same scale of test scores at *t*=2 – and the observed γ_1 with non-equated tests (see Section 3).

In the lower panel of Table A1 we compute DID according to equation (6):

$$DID = \Big(E(\gamma_2|I=1) - E(\gamma_2|I=0)\Big) - \Big(E(\gamma_1|I=1) - E(\gamma_1|I=0)\Big) \quad (6)$$

The true DID (based on equated test scores) is 23.5, whereas the observed DID (resulting from the differentiation of regression coefficients in the different metrics) is 3.5.

Table A1. Model parameters and DID with equated and observed scores

	Tracked	Untracked
ρ	50	30
β	40	20
θ	0.1	0.05
$\gamma_1\ (equated)$	50	30
$\gamma_1\ (observed)$	100	60
γ_2	95	51.5

$\omega = 0.5; \gamma_1\ (equated) = \rho;\ \gamma_1\ (observed) = \rho/\omega; \gamma_2 = \beta + (1+\theta)\rho$

$DID\ (equated) - true$	23.5
$DID\ (observed) - estimated$	3.5

$DID\ (equated)$ = (95-51.5)-(50-30)=23.5; $DID\ (observed)$ = (95-51.5)-(100-60)=3.5

In Table A2 we report the results of 10 simulations: after random generation of the data for each country and assessment according to the parameters described in Table A1, we estimate models M1 and M2 with two different sets of data for *t*=1: (i) with (latent) equated scores; (ii) with observed scores. It is clear that, for both models, the analyses with equated scores deliver results in line with the 'true DID' value and the analyses with observed scores are in line with the computed 'observed DID', hence, are heavily biased.

Table A2. Results of 10 replications of the simulation: DID with equated and observed scores

	M1		**M2**	
Replication	Equated scores	Observed scores	Equated scores	Observed scores
1	23.78	2.88	24.31	3.38
2	24.08	7.35	24.60	7.87
3	26.28	3.42	25.88	3.04
4	21.00	5.70	20.24	4.46
5	20.63	0.93	20.59	1.22
6	22.30	2.99	21.88	2.65
7	20.21	2.29	20.65	2.86
8	20.16	-1.71	20.23	-1.80
9	24.67	3.07	24.37	2.47
10	23.99	7.70	24.65	8.10

A.3 Difference-in-difference with within-countries standardized scores

Standardized variables are typically used to solve scaling problems and allow the direct comparison of quantities originally expressed in different metrics. In this perspective, we could take into consideration estimating individual pooled models of within-countries standardized scores. In this case, regression coefficients represent by how many SD the average achievement differs between pupils of different family backgrounds: thus, differentiating these quantities in not meaningless. Consider a positive *DID*. This result provides evidence that *for some reason*, the family background relative gap (relative to *each* country's SD) has increased more (or decreased less) in tracked countries than in untracked countries. As discussed in Section 3, this result could be due to different family background effects between t=1 and t=2 across institutional regimes, or instead to the different influence of other explanatory variables influencing achievement (even if nearly independent of family background, like gender). In other words, we could find a *DID* different from 0 even in the absence of any true effect of family background.

Appendix B. Empirical materials: variables' definitions and data sources

Table B.1 Variables' 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.
Gender	Female=0, Male=1

Table B.2 Age of tracking by country and data source

COUNTRY	AGE OF TRACKING	SOURCE
Austria	10	Eurydice: "The structure of European Education systems 2012", European Commission
Belgium	14	Eurydice: "The structure of European Education systems 2012", European Commission
Bulgaria	14	Eurydice: "The structure of European Education systems 2012", European Commission
Canada	18	Education system Canada-EP Nuffic (2015) "The Canadian system described and compared with the Dutch system"
Denmark	16	Eurydice: "The structure of European Education systems 2012", European Commission
France	15	Eurydice: "The structure of European Education systems 2012", European Commission
Germany	10	Eurydice: "The structure of European Education systems 2012", European Commission
Hungary	10	Eurydice: "The structure of European Education systems 2012", European Commission
Israel	15	Education system Israel-EP Nuffic (2015) "The Israeli system described and compared with the Dutch system"
Italy	14	Eurydice: "The structure of European Education systems 2012", European Commission
Latvia	16	http://www.aic.lv/portal/en/izglitiba-latvija
Lithuania	16	http://education.stateuniversity.com/pages/872/Lithuania-EDUCATIONAL-SYSTEM-OVERVIEW.html
Luxembourg	12	Eurydice: "The structure of European Education systems 2012", European Commission
Netherlands	12	Eurydice: "The structure of European Education systems 2012", European Commission
New Zealand	16	http://www.oecd.org/education/EDUCATION%20POLICY%20OUTLOOK%20NEW%20ZEALAND_EN.pdf
Norway	16	Eurydice: "The structure of European Education systems 2012", European Commission
Poland	16	Eurydice: "The structure of European Education systems 2012", European Commission
Romania	16	Eurydice: "The structure of European Education systems 2012", European Commission Eurydice Italia: "Sistemi scolastici europei 2012", Indire, European Commission.
Russian Fed.	16	http://www.alberta.ca/documents/IQAS/russia-international-education-guide.pdf
Slovakia	11	OECD (2014): "OECD Reviews of Evaluation and Assessment in Education Slovak Republic"
Slovenia	15	Eurydice: "The structure of European Education systems 2012", European Commission
Spain	16	Eurydice: "The structure of European Education systems 2012", European Commission
Sweden	16	Eurydice: "The structure of European Education systems 2012", European Commission
USA	18	https://isss.umn.edu/publications/USEducation/2.pdf

Appendix C. Second-step results. Robustness checks.

Table C.1. Countries with tracking at age 15 classified as untracked.

	INEQUALITY MEASURE			
	Overall inequalities		Family background inequalities	
	SD		**F-GAP**	
	(1)	(2)	(3)	(4)
Constant	58.17***	65.24***	66.31***	88.77***
Tracking	6.50*	-15.15	20.42***	-34.11
Inequality measure at *t*=1	0.486***	0.387**	0.603***	0.336*
Tracking* Inequality measure at *t*=1		0.328		0.657**
N countries	24	24	24	24
R^2	*0.368*	*0.400*	*0.558*	*0.657*

NOTES. Under the heading SD we report results relative to the effects of tracking on the country standard deviation. Under the heading F-GAP we report results relative to the effects of tracking on the difference between tertiary educated parents with the largest number of books and no tertiary educated parents with the lowest number of books: F-GAP=[ln(500)*β(*ln*(n° books))+β(tertiary degree)]- ln(5)*β(*ln*(n° books).
Columns (1) and (3) refer to the model with no interaction; columns (2) and (4) to models with interaction between the tracking indicator and inequality at *t*=1.
* 0.05<p-value<0.10, ** 0.01<p-value<0.05, *** p-value<0.01

Table C.2. Tracking variable: Number of years since tracking

	INEQUALITY MEASURE			
	Family background inequalities		Overall inequalities	
	F-GAP		**SD**	
	(1)	(2)	(3)	(4)
Constant	76.71***	82.05***	63.32***	63.60***
N° years since tracking	4.00**	-3.86	0.848	0.494
Inequality measure at *t*=1	0.538***	0.473***	0.438***	0.434**
N° years since tracking* Inequality measure at *t*=1		0.089		0.004
N countries	24	24	24	24
R^2	*0.503*	*0.540*	*0.286*	*0.287*

NOTES. N° years since tracking is defined as (15- age of tracking) if tracking occurs up to age 15, is equal to -1 if tracking occurs after age 15 (not yet occurred). Under the heading SD we report results relative to the effects of tracking on the country standard deviation. Under the heading F-GAP we report results relative to the effects of tracking on the difference between tertiary educated parents with the largest number of books and no tertiary educated parents with the lowest number of books: F-GAP=[ln(500)*β(*ln*(n° books))+β(tertiary degree)]- ln(5)*β(*ln*(n° books).
Columns (1) and (3) refer to the model with no interaction; columns (2) and (4) to models with interaction between the tracking indicator and inequality at *t*=1.
* 0.05<p-value<0.10, ** 0.01<p-value<0.05, *** p-value<0.01

Supplementary material

First step estimates

Table 1. Scores standard deviations at *t*=1 (PIRLS) and *t*=2 (PISA)

	Primary school PIRLS (2006)		Secondary school PISA (2012)	
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 2. Regression estimates. 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
Lithuania	497.23	-16.21	4.81	11.72	6.57	9.53	31.2	17.4
se	3.69	2.27	2.42	2.79	3.19	0.77	2.46	1.1
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 3. Regression estimates. 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
Lithuania	411.77	-47.75	4.82	11.8	16.15	19.77	16.45	24.79
	4.63	2.19	2.87	3.16	3.73	0.983	3.07	1.25
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.

Additional descriptive analyses on first step results.

In Table 4 we examine some correlations of the first step estimates of SD and F-GAP across countries providing some insights on the relation between *overall inequalities* (measured by the test scores' standard deviation) and *family background inequalities* (measured by a linear combination of the regression coefficients of the two social background variables: log number of books and parent with tertiary education). We consider statistics based on absolute values and statistics relative to country rankings (1 is the smallest value and N the highest value).

Inequalities over time

Countries displaying larger inequality in primary school also tend to display larger inequalities in secondary school (columns 1-2). Correlations between social background differentials (F-GAP) at t=1 and t=2 are stronger than the correlation between SDs, and substantially larger within tracked countries than within untracked countries. Interestingly, the correlation coefficient between ΔSD and ΔF-GAP 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. This tells us that countries raising their relative position with respect to social background inequality also tend to raise their relative position with respect to overall inequality. (Notice that we do not compute the correlation on original scores, because, as we have seen in the section 3, the difference of regression coefficients at different assessments ΔF-GAP has no substantive meaning, and the same argument would apply to ΔSD).

Overall inequalities and social background inequalities

As shown in columns 3-4, the cross-country correlation between SD and F-GAP is 0.617 at t=1 and 0.659 at t=2. If we consider country rankings instead of absolute figures, we obtain 0.667 at t=1 and 0.578 at t=2, confirming that overall inequalities and family background inequalities are related, but their relation need not to be very strong. The reason is that overall inequalities are driven by family background educational inequalities, but may also depend on other factors.

Table 4. Cross-country correlations on absolute measures and rankings

ABSOLUTE MEASURES					
	(1)	(2)	(3)	(4)	(5)
	SD1, SD2	F-GAP1, F-GAP2	SD1, F-GAP1	SD2, F-GAP2	ΔSD, ΔF-GAP
Tracked	0.760	0.780	0.517	0.538	-
Untracked	0.263	0.563	0.770	0.663	-
All	0.492	0.569	0.617	0.659	-
RANKINGS					
	SD1, SD2	F-GAP1, F-GAP2	SD1, F-GAP1	SD2, F-GAP2	ΔSD, ΔF-GAP
Tracked	0.647	0.818	0.756	0.363	0.814
Untracked	0.205	0.743	0.613	0.534	0.498
All	0.323	0.688	0.667	0.578	0.738

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

F-GAP describes the gap between tertiary educated parents with the largest number of books and no tertiary educated parents with the lowest number of books: F-GAP=[ln(500)*β(*ln*(n° books))+β(tertiary degree)]- ln(5)*β(*ln*(n° books)). Model errors correlated within countries.