Contini and Cugnata *Large-scale Assess Educ* (2020) 8:14
https://doi.org/10.1186/s40536-020-00094-x

Large-scale Assessments in Education

RESEARCH Open Access

# Does early tracking affect learning inequalities? Revisiting difference-in-differences modeling strategies with international assessments

Dalit Contini[1*] and Federica Cugnata[2]

*Correspondence:
dalit.contini@unito.it
[1] Department of Economics and Statistics, University of Torino, Turin, Italy
Full list of author information is available at the end of the article

**Abstract**

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. In this contribution, we discuss the difference-in-differences strategies employed in the existing literature to evaluate the effect of early tracking on learning inequalities exploiting international assessments administered at different age/grades. In their seminal paper, Hanushek and Woessmann (Econ J 116:C63–C76, 2006) analyze with two-step estimation the effect of early tracking on *overall* inequalities, measured by test scores' variability indexes. Later work of other scholars in the economics and sociology of education focuses instead on inequalities among children of different *family background*, using individual-level models on pooled data from different countries and assessments. In this contribution, we show that individual pooled difference-in-differences 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. Starting from a simple learning growth model, we show that if test scores at different surveys are not measured on the same scale, as occurs for international learning assessments, pooled individual models may deliver severely biased results. Instead, the scaling problem does not affect the two-step approach. For this reason, we suggest using two-step estimation also to analyze *family-background* achievement inequalities. Against this background, using PIRLS-2006 and PISA-2012 we conduct two-step analyses, finding new evidence that early tracking fosters both overall inequalities and family background differentials in reading literacy.

**Keywords:** International assessments, Test scores, Achievement inequalities, Cross-country analyses, Educational systems, Early tracking, Difference-in-differences

Springer Open

© The Author(s) 2020. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## 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 et al. 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; 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 and Pop-Eleches 2011 on Romania;

Piopiunik 2014 on Bavaria; Kerr et al. 2013 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 a *difference-in-differences* strategy. This method, commonly used in econometric analyses, solves the problem of unobserved country-level heterogeneity by analyzing the outcome variable at two or more time points and examining the extent to which differences over time vary between treated and control units (hence, difference-in-differences). The underlying assumption is that if no causal effect were at work the two groups would experience the same time change: evidence of a divergence in this change signals a treatment effect. In particular, Hanushek and Woessmann (2006) apply difference-in-differences 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-differences to *family-background* differentials (e.g. Waldinger 2007; Jakubowski 2010; Ammermueller 2013; Ruhose and 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 analyze variability at $t=2$ as a function of variability at $t=1$ and the early tracking indicator. As described below in more detail, the parameter of interest is the coefficient of early tracking in this second step. The later studies, instead, pool together the micro-data from all countries and both assessments, and estimate individual-level achievement models with individual- and system-level explanatory variables (including the tracking regime), the time of the survey, various two-way interaction terms and a three-way interaction term between family background, time of the survey and the early tracking indicator. The coefficient of the latter term is meant to capture the effect of tracking on family background inequalities (i.e. the difference between the variation in the family background coefficients over time in early and late tracking regimes).

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 Woessmann (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 the next section, 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 Woessamnn (2006), who analyze the effect of early tracking on country-level variability measures, and then move to the individual pooled models used to study the 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 (cross-sectional) family background regression coefficients between tracking regimes and learning assessments. In the next section, 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. We then analyze how the scaling issue affects the results of individual pooled difference-in-difference models and demonstrate that the estimates of institutional effects delivered by pooled individual models may be severely biased. In the following section, extending the simple approach of Hanushek and Woessamnn (2006) to the analysis of the effect of early tracking on family background inequalities, we propose a more flexible two-step estimation strategy, first describing individual achievement differentials within countries and then relating family-background regression coefficients to institutional variables. In the next section, we describe our empirical analysis and discuss the results. Conclusions follow.

## Literature review

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 (Woessmann 2005, 2010; Fuchs and Woessmann 2007), 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 Woessmann (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 Woessmann (2006) use PIRLS (4th grade) + PISA (age 15) to investigate the effects of

tracking on reading literacy and TIMSS (4th grade) + TIMSS (8th grade) to investigate the effects on math. The rationale is that while in 4th grade children are still in comprehensive school everywhere, in 8th 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) 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-differences to study the effect of early tracking on achievement inequalities related to the migration background. Differently from Hanushek and Woessmann (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.

Before moving to the 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* focuses 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* focuses 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 $\gamma$ be the family background coefficient at a given survey. In a stylized 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 $\sigma_y^2$ depends on the family-background-specific effect ($\gamma$), on the variability of family background in the population ($\sigma_x^2$), and on the influence of other factors independent of family background ($\sigma_\varepsilon^2$). 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.

#### Overall inequalities: Hanushek and Woessmann's seminal paper

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 $I$ indexing early tracking, given the *SD* at $t=1$. In particular, they estimate the simple linear model:

$$SD_{2c} = a + bSD_{1c} + dI_c + u_c \tag{1}$$

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. Indeed, (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.

#### 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.[1]

The simplest model adopted in the literature (Waldinger 2007; Jakubowski 2010; Ruhose and Schwerdt 2016) is:

Model M1

$$Y_{itc} = \alpha_{0c} + \alpha_1 T + \alpha_2 I_c T + \pi X_{itc} + \xi_1 F_{itc} + \lambda_1 F_{itc} I_c + \xi_2 F_{itc} T + \lambda_2 F_{itc} T I_c + \varepsilon_{itc} \tag{2}$$

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,

[1] 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.

$T$ 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 $\alpha_{0c}$ is a country-specific fixed effect (estimated with country dummy variables), thus it need not to be independent of the other explanatory variables. The parameter of main interest is $\lambda_2$, the coefficient of the 3-level interaction term. Denote the family background coefficients at $t=1$ and $t=2$ as $\gamma_1$ and $\gamma_2$. The following relations hold: $\gamma_1 = \xi_1 + \lambda_1 I$, and $\gamma_2 = (\xi_1 + \xi_2) + (\lambda_2 + \lambda_2)I$. 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.[2]

A more flexible model includes country/time fixed effects (Ammermueller 2013) is:

Model M2

$$Y_{itc} = \alpha_{0tc} + \pi X_{itc} + \xi_{1c} F_{itc} + \xi_2 F_{itc} T + \lambda_2 F_{itc} T I_c + \varepsilon_{itc} \tag{3}$$

Here the intercept $\alpha_{0tc}$ 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 $\gamma_{2c} = \xi_{1c} + (\xi_2 + \lambda_2 I)$. 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 $I$.

Since in these models the relation between $\gamma_2$ an $\gamma_1$ can be expressed as:

$$\gamma_{2c} = \gamma_{1c} + (\xi_2 + \lambda_2 I_c) \tag{4}$$

(in M1 this further simplifies, since $\gamma_{1c} = \xi_1 + \lambda_1 I_c$), the parameter of main interest $\lambda_2$ corresponds to the standard difference-in-difference definition:

$$DID = \big(E(\gamma_2 | I = 1) - E(\gamma_2 | I = 0)\big) - \big(E(\gamma_1 | I = 1) - E(\gamma_1 | I = 0)\big) \tag{5}$$

representing the double difference in the family background regression coefficients between the two surveys (i.e. between secondary and primary schooling), and between

[2] 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).

tracked ($I=1$) and untracked ($I=0$) educational systems, but can also be interpreted as $E(\gamma_2|\gamma_1, I=1) - E(\gamma_2|\gamma_1, I=0)$.

## Validity of pooled individual difference-in-differences models

In this section we discuss the validity of the results delivered by pooled individual difference-in-differences models when evaluating the effect of institutional characteristics on learning inequalities. First, we review the scaling issue when comparing the achievement of children in different assessments, second we focus on the consequences on the difference-in-differences models employed in the existing literature.

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 educational 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[3] 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 will 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 rationale is the following. If expressed in different metrics, cross-sectional regression coefficients are not comparable across surveys: their difference $(\gamma_2 - \gamma_1)$ is meaningless. We show this rather trivial point below, based on a stylized structural achievement growth model. We will then re-interpret under these lenses the results delivered by the difference-in-difference strategies based on individual pooled models. For some reason, the scaling issue has been ignored in this literature: we presume that the implicit assumption is that with the double differencing the scaling issue would disappear. We will show this is generally not the case.

### Comparing achievement inequalities as children grow: the scaling issue

To analyze the evolution of inequalities at different stages of schooling, we have to compare test scores' inequality measures across assessments administered to children of different age. 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 test scores delivered by the international agencies (standardized across countries), and conveys only limited information on the evolution of inequalities when using within-country standardized

[3] International Association for the Evaluation of Educational Achievement.

test scores (produced by standardizing original scores relative to each country's mean and SD).

Consider a stylized model of learning development according to which abilities of a cohort of children 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.

#### *Same scale*

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:

$$y_{i1} = \mu_1 + \rho F_i + \varepsilon_{i1}. \tag{6}$$

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 $\delta$:

$$y_{i2} = y_{i1} + \delta_i. \tag{7}$$

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

$$\delta_i = \Delta + \beta F_i + \theta y_{i1} + \varepsilon_{i2} \tag{8}$$

$\beta$ measures whether children of high backgrounds improve or worsen their performance between $t=1$ and $t=2$, relative to equally performing children of low backgrounds at $t=1$: we call this "new inequalities" developed between the two assessments. Instead, $\theta$ captures carry-over effects of pre-existing inequalities. The total effect of family background operating between $t=1$ and $t=2$ is $(\beta + \rho\theta)$, given by the sum of the direct effect given previous achievement and the indirect effect via previous achievement.

With longitudinal data it is possible to evaluate achievement growth for each child, estimate model (8) and identify the structural parameters $\beta$ and $\theta$, disentangling the two different mechanisms at play in the development of inequalities over time. With cross-sectional data, the longitudinal model and the structural parameters model $\beta$ and $\theta$ are not identified.[4] Nonetheless, the total effect $(\beta + \rho\theta)$ is still identified. Consider the cross-sectional model at $t=2$:

[4] In particular circumstances identification of $\beta$ is possible with pseudo-panel techniques (Contini and Grand 2017).

$$y_{i2} = \Delta + \beta F_i + (1+\theta)y_{i1} + \varepsilon_{i2} = cost + (\beta + (1+\theta)\rho)F_i + (1+\theta)\varepsilon_{i1} + \varepsilon_{i2}.$$

The cross-sectional coefficient at $t=2$ is:

$$(\beta + (1+\theta)\rho) \tag{9}$$

and the difference between the coefficients at $t=1$ and $t=2$ is $(\beta + (1+\theta)\rho) - \rho = \beta + \rho\theta$.

#### *Different scales*

International learning assessments, as many national surveys, are cross-sectional, and achievement is measured on different scales as children grow older. Moreover, test scores are not vertically equated. In this case, we have to distinguish between the observed scores $y'_1$ and the (unknown) scores $y_1$ representing achievement at $t=1$ according to the scale employed at $t=2$. In this case, the difference between the cross-sectional regression coefficients at $t=2$ and $t=1$ does not identify $\beta + \rho\theta$. Assuming for simplicity a linear relation between these scales (where $\varphi$, and $\omega$ are unknown and unidentifiable):

$$y_{i1} = \varphi + \omega y'_{i1} \tag{10}$$

from (6) we obtain the model relating observed scores $y'_{i1}$ to family background $F$:

$$y'_{i1} = const + \frac{\rho}{\omega}F_i + \frac{\varepsilon_{i1}}{\omega} \tag{11}$$

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. The coefficient at $t=2$ is given by (9). Patently, if $\omega \neq 1$, the difference between the estimable cross-sectional regression coefficients at the two assessments:

$$\beta + (1+\theta)\rho) - \frac{\rho}{\omega} \tag{12}$$

differs from $\beta + \rho\theta$ and 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). If we want to analyze the development of family background inequalities in a given country, the standardization is obtained relative to the country mean and standard deviation. In a regression framework, this amounts to comparing regression coefficients of models run on within-country standardized scores. Being invariant to the score metric, these quantities are comparable:

$$E(z_1|F=1) - E(z_1|F=0) = \frac{\rho/\omega}{\sigma_{y_1'}} = \frac{\rho}{\sigma_{y_1}} \tag{13}$$

$$E(z_2|F=1) - E(z_2|F=0) = \frac{(1+\theta)\rho + \beta}{\sigma_{y_2}} \tag{14}$$

The difference between (13) and (14) informs on how many standard deviations two individuals of different family backgrounds are apart at $t=2$ as compared to $t=1$. However, within-country variability is generally not the same across assessments, so this difference also depends on the standard deviations. As a result, the sources of the observed change are unclear. Children's achievement is not influenced only by family background: if in a country 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 $\beta > 0$.[5] Hence, even if the comparison of regression coefficients on standardized scores is not meaningless, their difference does not allow to identify $\beta + \rho\theta$ and is not fully informative on how inequalities related to family background evolve over time.

#### *Relating cross-sectional regression coefficients at different surveys*

Once again, let us denote the family background coefficients at $t=1$ and $t=2$ as $\gamma_1$ and $\gamma_2$. Under the stylized achievement growth model (6)–(10), it is trivial to show that with *same scale* scores the relation between the regression coefficients at the two cross-sectional assessments is:

$$\gamma_2 = \gamma_1 + k \tag{15}$$

because $\gamma_1 = \rho$ and $\gamma_2 = (\beta + (1+\theta)\rho) = \rho + (\beta + \rho\theta)$, so $k = \beta + \rho\theta$.

Instead, with *different scale* test scores:

$$\gamma_2 = \omega\gamma_1 + k \tag{16}$$

because $\gamma_2 = \rho + (\beta + \rho\theta)$ and $\gamma_1 = \frac{\rho}{\omega}$, $\gamma_2 = \omega\frac{\rho}{\omega} + (\beta + \rho\theta) = \omega\gamma_1 + k$.

This result is crucial because the implied relation (4) existing between the $F$-coefficients in pooled individual difference-in-differences models M1 and M2, corresponds to the case where $\omega = 1$, thus is not valid under the different-scale case, occurring for international learning assessments.

### The scaling issue in difference-in-differences pooled individual models

We have shown above that the difference between regression coefficients $\gamma_2$ and $\gamma_1$ when test scores are not equated is generally meaningless. In the previous section we reviewed the difference-in-difference strategies employed in the literature on educational inequalities and highlighted that, in essence, individual pooled models identify the effect of early tracking or other institutional features on family background inequalities, by taking

[5] However, it can be shown however that a positive difference between (13) and (14) implies $\beta > 0$.

the (double) difference of cross-sectional regression coefficients relative to assessments administered at different children's age.

The key question now is: *Does the double differentiation of regression coefficients solve the scaling problem?* Starting from the stylized achievement growth model presented above, we now show that the answer is no.

To fix ideas, for a specific cohort of children think of PIRLS (4th grade) as the assessment at $t=1$ and PISA (age 15) as the assessment at $t=2$.[6] 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 (9) and (11). Thus, regression coefficients, in the most general setting variable across countries, may be expressed as:

$$\begin{aligned} &\gamma_{1c} = \rho_c/\omega \text{ at } t=1 \\ &\gamma_{2c} = \beta_c + (1+\theta_c)\rho_c \text{ at } t=2 \end{aligned} \tag{17}$$

where $\omega$ reflects the different scale used to measure test scores in the two surveys.

#### *Difference-in-differences with model M1*

In model M1, regression coefficients are allowed to vary across countries only according to the tracking system. Recall that in this case DID amounts to:

$$(E(\gamma_2|I=1) - E(\gamma_2|I=0)) - (E(\gamma_1|I=1) - E(\gamma_1|I=0))$$

where $\gamma_1$ and $\gamma_2$ are the cross-sectional regression coefficients of family background in the two assessments, $I=1$ represents early tracking systems and $I=0$ late tracking systems. Substituting the structural parameters (17) into this expression, and recalling that M1 assumes that the coefficients vary across countries only according to the tracking regime, we obtain:

$$DID = [(\beta_{I=1} + (1+\theta_{I=1})\rho_{I=1}) - (\beta_{I=0} + (1+\theta_{I=0})\rho_{I=0})] - [(\rho_{I=1} - \rho_{I=0})/\omega] \tag{18}$$

The first term in square brackets is the difference between the regression coefficients in tracked and untracked systems in secondary school; the second one is the difference between the regression coefficients in tracked and untracked systems in primary school. This expression 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 ($\rho_{I=1} = \rho_{I=0}$); (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 $\beta_{I=1} + (1+\theta_{I=1})\rho_{I=1} = \beta_{I=0} + (1+\theta_{I=0})\rho_{I=0}$. 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.[7]

[6] As the (longitudinal) growth model (7), (8) refers to a specific population of children, the cross-sectional coefficients derived in the paper apply to children belonging to the same cohort, or conditional on the assumption that no cohort-effects exist.

[7] See also the simulation exercise in the Additional file 1, Appendix.

***Difference-in-differences with model M2***

In model M2 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 $\rho$ vary freely across countries (indicated as $\rho_c$), but constrain $\beta$ and $\theta$ to depend on tracking. Substituting the corresponding regression coefficients into the expression for the standard *DID* we obtain:

$$[(\beta_{I=1}+(1+\theta_{I=1})E_{I=1}(\rho_c))-(\beta_{I=0}+(1+\theta_{I=0})E_{I=0}(\rho_c))]-[E_{I=1}(\rho_c)-E_{I=0}(\rho_c)]/\omega \tag{19}$$

where $E(\rho_c)$ is the expected value of $\rho$ in a given tracking regime. Once again, the estimated *DID* depends on the unknown scaling factor $\omega$ and delivers meaningful results only under the fortuitous circumstances described above for M1.[8]

A final consideration is in order. To illustrate that double differencing does not solve the scaling problem we have relied on the restrictive stylized achievement growth model, but we believe our conclusions are far more general. In essence, we have shown that the flaws of individual pooled models M1–M2 are due the fact that they implicitly assume same scaling and that double differencing does not help: if these considerations apply to a simple model, they are very likely to hold also under more complex ones.

## Two-step estimation

We have shown that under the stylized achievement growth model described above, the relation between the *F*-regression coefficients is of the type: $\gamma_2=\omega\gamma_1+k$, but M1 and M2 implicitly impose $\omega=1$. One might consider a more flexible two-level model—let us call it M3—not imposing this restriction, with an individual-level model for each country and assessment, and a country-level model relating regression coefficients and institutional characteristics. In this sense, this model could be conceived as a "generalized" difference-in-differences model. An additional advantage of this specification is its transparency: first- and second- step models are simple, their underlying assumptions are clear, and the interpretation of the results is straightforward.

Model M3

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:

$$Y_{itc}=\alpha_{0tc}+\gamma_{tc}F_{itc}+\varphi_{tc}X_{itc}+\varepsilon_{itc} \tag{20}$$

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:

$$\gamma_{2c}=a+b\gamma_{1c}+dI_c+u_c \tag{21a}$$

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

[8] See also the simulation exercise in the Additional file 1, Appendix.

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:

$$\gamma_{2c} = a + b\gamma_{1c} + dI_c + g\gamma_{1c}I_c + u_c \tag{21b}$$

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 is consistent with the non-standard *DID* definition:

$$DID = E(\gamma_2|\gamma_1, I = 1) - E(\gamma_2|\gamma_1, I = 0)$$

previously employed in Hanushek and Woessmann (2006) to analyze the effect of early tracking of overall inequalities, conceived as test scores' variability. 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 $\gamma_{1c}$ and a substantial overlap of the distributions of $\gamma_{1c}$ between the subgroups of countries identified by $I=0$ and $I=1$.[9]

Estimation of model M3 can be carried out in two steps, as in Hanushek and Woessmann (2006).

*Step 1* In the first step, the family-background regression coefficients in (20) 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 (21a) or (21b). 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 (21a), (21b) is the existence of sufficient variability in the $\gamma_{1c}$ distributions within institutional regimes.[10] 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

[9] Cross-country variability in $\gamma_1$ is necessary for the model identification. A substantial overlap of the distributions of $\gamma_1$ between the subgroups of countries is necessary because we are aiming at estimating inequalities at $t=2$ *given* inequalities at $t=1$, thus *at the same level* of $\gamma_1$.

[10] Using estimated values from a previous stage as a dependent variable in a second stage introduces downward biased standard errors of the coefficients' estimates, because the second step ignores the estimation error from the first stage. Different software programs provide routines to address this specific issue. In the present context, however, not only the dependent variable is estimated in a previous stage, but also an independent variable (the $F$-regression coefficient at $t=1$). This should lead to bias in the effect of the treatment variable $I$. At present, we are not aware of simple solutions to this problem, so we neglect this issue. However, due to the large sample sizes in the first step (in the countries of interest in PIRLS 2006: N = 3500–8000 and in PISA 2012: N = 5000–38,000), measurement error should be small and it should not lead to substantial bias in second stage estimation.

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

#### Difference-in-differences definitions

Even if model M3 is more general than M1 and M2, the conclusion that when test scores are not on the same scale the standard $DID = (E(\gamma_2|I=1) - E(\gamma_2|I=0)) - (E(\gamma_1|I=1) - E(\gamma_1|I=0))$ delivers meaningless results holds true also for M3.

The advantage of M3 is that here the identification of the effect of early tracking is reached by estimating the non-standard *DID*: $E(\gamma_2|\gamma_1, I=1) - E(\gamma_2|\gamma_1, I=0)$, in second step models (21a) or (21b), that directly relate regression coefficients at $t=2$ to the tracking regime and regression coefficients at $t=1$. The different scaling is no longer a problem because in regression models there is no need for dependent and independent variables to be on the same scale (unless a priori restrictions on the coefficient of dependent variables are imposed, as implicit in M2).

#### Identification of the structural parameters?

While the results described so far are very general, under the strict assumptions of the stylized achievement growth model, we could even take some steps further and derive some conclusions on the mechanism at play. According to (17), the following holds:

$$E(\gamma_{2c}|\gamma_{1c}) = E(\beta_c) + \omega(1 + E(\theta_c))\gamma_{1c}. \tag{22}$$

Thus $\beta$ is related to the intercept and $\theta$ to the slope. Let us allow $\gamma_{2c}$ to vary with the tracking regime, according to (21a) and (21b):

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

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

Equation (22) is consistent with the first specification if on average $\beta$ (new family background inequality developed between $t=1$ and $t=2$) varies across countries with the tracking regime and $\theta$ (carry-over effect of previously established inequalities) does not vary with the tracking regime. It is consistent with the second if both $\beta$ and $\theta$ vary with the tracking regime.

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

Due to the dependence of this result on restrictive assumptions, caution is advised when interpreting two-step results in this manner, as the linear specification may be only a convenience approximation of a potentially more complex relation between previous and later achievement gaps. In addition, the intercept's estimate is usually unstable with

**Table 1 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 |

Source: see Appendix A, Table 5

Dummy tracking: = 1 if tracking occurs at age $\leq$ 15, 0 otherwise

Alternative definition (see Robustness checks in Appendix B): Dummy tracking: = 1 if tracking at age $\leq$ 14, 0 otherwise

small sample size, as occurs with cross-country models relying on a limited number of countries.

## Empirical analysis

Based on the methodological considerations developed in the previous sections, we now illustrate the analysis of the effect of early tracking on family background inequalities with two-step estimation, 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. Moreover, as highlighted in Jakubowski (2010), some migrants were not in the country of the test in the 4th grade or were not fully exposed to the country schooling system, so the results from the two assessments may not be fully comparable. 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.

In line with the methodological considerations developed in the previous sections, we apply two-step analysis to *family background inequalities*, but we also analyze *overall inequalities* (replicating the analyses carried out by Hanushek and Woessmann (2006) on more recent data and a different set of countries). In the first step, for each country and assessment we estimate the test scores standard deviations and the family background regression coefficients with model (20). We include two variables to measure family background, related to cultural capital: the log-number of books, regarded in the literature as the best single predictor of student performance (Hanushek and Woessmann 2011), and a binary variable indexing whether at least one parent has tertiary education.[11] We also control for gender and age (see Appendix A 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 $t=1$.

#### 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.[12]

To capture the effect of tracking on family background inequalities, instead of looking separately at the two explanatory variables, we focus on the linear combination of the coefficients of the two variables log-number of books and the parental education dummy, highlighting the effect of tracking on the test-scores differential between children with tertiary educated parents and "many" books (n = 500), and children with non-tertiary educated parents and "few" books (n = 5) books. If $c_1$ and $c_2$ are the estimates of the two coefficients, in the tables below we report $c_1(\ln(500) - \ln(5)) + c_2$ and name it F-gap.

[11] We use the number of books in the home and parental education (as reported by children) as measures of SES as they are the most frequently employed in this strand of literature. This occurs most likely because: i) as already stated, the number of books is the single best predictor of achievement; ii) they are available in both assessments and are flawed by lower shares of missing data than the corresponding parental reports, or than information on parental occupation. It must be noted, however, that based on comparisons between children's and parents' reports, and assuming that the latter are correct, these measures do not appear to have high reliability (Jerrim and Micklewright 2014). In the presence of classical measurement error, the consequence would be the underestimate of the SES effect but if the errors are similar in size across countries, the country rankings of inequality should not be affected (ibid.). The direction of the bias is difficult to predict in difference-in-differences analyses, because it applies to both an independent and the dependent variables (see also footnote 10, dealing with another source of possible measurement error). This issue is out of the scope of the present paper, so we will not make further reference to it, and assume that our main results will not be heavily affected by measurement error in SES indicators.

[12] The full set of first step results is available from the authors upon request.

**Table 2 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-gap 2 |
| 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) |

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: $F\text{-}_{GAP} = [\ln(500) \times c_1 + c_2] - \ln(5) \times c_1$

**Table 3 Second step results. Cross-country regression models**

| | Inequality measure | | | |
|---|---|---|---|---|
| | Overall inequalities | | Family 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 |

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: $\text{F-gap} = [\ln(500) \times c_1 + c_2] - \ln(5) \times c_1$

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

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 2). A similar pattern holds when looking at family 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.

### 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 *family background inequalities*, we estimate models (21a) and (21b) 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 3. The coefficients of prior inequalities are always

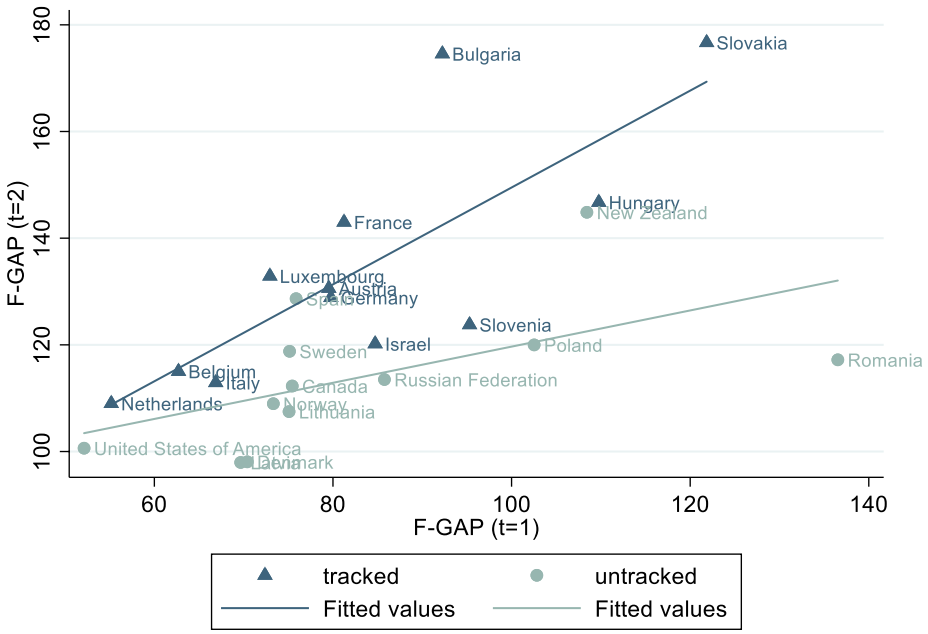


**Fig. 1** Family background effects at $t=2$ and $t=1$. F-gap in primary (PIRLS) and secondary school (PISA)

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 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 family background inequalities—columns (3) and (4)—indicate that early tracking is associated to larger inequalities. Given educational inequality already existing in primary school, the F-gap at age 15 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 B, Table 6), whereas no interaction effect is observed when considering the number of years since tracking (Appendix B, Table 7).

In Fig. 1 we show the scatter diagram depicting observed family 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 (21b) reported in column (4) of Table 3. First of all, this graph shows that in primary school family background inequalities vary considerably across countries even within tracking regimes. Secondly, it shows that at low levels of family 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.

For illustrative purposes, we now attempt to interpret the results relative to the effect of tracking on family background inequalities in terms of the structural parameters of the achievement growth model ("Validity of pooled individual difference-in-differences models"). As already remarked, however, due to the restrictive underlying assumptions and the small sample size in the step-2 estimation, this structural interpretation of the results is to be taken with caution.

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 $\beta$, are similar in tracked and untracked systems (or perhaps even lower in tracked systems, since the point estimate is negative although not statistically significant). Instead, carry-over effects of previous family background inequalities in achievement, represented by $\theta$, seem to be 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 family 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 attend the academic track as compared to more labor-market oriented schools more often than low-performing children, although with different probabilities across family backgrounds. Thus, the gap between well- and poor- performing children may widen more sharply in these countries than in comprehensive school systems.

#### 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 1. 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 the value 20.40 delivered by the two-step estimation model (21a) and showed in Table 3 column (3). 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 2, the mean $F$-gap is 83.5 points in tracked countries and 83.4 in untracked countries. Thus, in this particular case we fall into one of the *fortuitous* circumstances thoroughly discussed in "The scaling issue in difference-in-differences pooled individual models", where the results of M1 and M2—although delivered by unnecessarily restrictive and weakly transparent models—*happen* not to be meaningless, as the estimated *DID* ends up being expressed in the metric of test scores at $t=2$.

## Discussion and conclusions

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-differences 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-differences 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 difference-in-differences pooled individual level models. Scaling issues do not apply instead to two-step estimation. Hence, provided that difference-in-differences 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 institutional effects[13] and different research areas. For example, 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 family 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 family 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 family 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

[13] For example, strength of the private sector, the degree of autonomy and time devoted to instruction.

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.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1186/s40536-020-00094-x.

**Additional file 1:** Additional material.

**Author details**
[1] Department of Economics and Statistics, University of Torino, Turin, Italy. [2] Vita-Salute San Raffaele University, Milan, Italy.

## Appendix A: Empirical materials: variables' definitions and data sources

See Tables 4 and 5.

**Table 4 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)<br>Children report the number of books at home, based on pictures depicting different numbers of shelves<br>Classification in PIRLS is 0–10; 11–25; 26–100; 101–200, > 200<br>Classification in PISA is 0–10; 11–25; 26–100; 101–200, 201–500, > 500<br>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)<br>In practice we use the following values:<br>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<br>No parents with tertiary education = 0 |
| Control variables | |
| Age | Country-specific quartiles' dummy variables (1°–4°)<br>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<br>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 5 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 | https://www.aic.lv/portal/en/izglitiba-latvija |
| Lithuania | 16 | https://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 | https://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<br>Eurydice Italia: "Sistemi scolastici europei 2012", Indire, European Commission |
| Russian Fed | 16 | https://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 B: Second-step results. Robustness checks

See Tables 6 and 7.

**Table 6 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 |

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 = $[\ln(500) \times c_1 + c_2] - \ln(5) \times c_1$, where $c_1$ and $c_2$ are the coefficients of the family background variables: ln(books) and tertiary education

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 7 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*** |
| No years since tracking | 4.00** | − 3.86 | 0.848 | 0.494 |
| Inequality measure at $t=1$ | 0.538*** | 0.473*** | 0.438*** | 0.434** |
| No 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 |

No 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 = $[\ln(500) \times c_1 + c_2] - \ln(5) \times c_1$, where $c_1$ and $c_2$ are the coefficients of the family background variables: ln(books) and tertiary education

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

Received: 21 March 2020 Accepted: 6 October 2020
Published online: 21 November 2020

**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., Diamond, A., & Heinmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science, 59*(2), 495–510.
Bol, T., Witschge, J., Van de Werfhorst, H. G., & Dronkers, J. (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., & Lang, K. (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., & Contini, D. (2014). Migrant achievement penalties in Western Europe. Do educational systems matter? *European Sociological Review, 30*(5), 670–683.
Brunello, G., & Checchi, D. (2007). Does school tracking affect equality of opportunity? New international evidence. *Economic Policy, 52,* 781–861.
Bryan, M. L., & Jenkins, S. P. (2016). Multilevel modelling of country effects: a cautionary tale. *European Sociological Review, 32*(1), 3–22.
Caro, D. H., & Biecek, P. (2017). intsvy: An R package for analyzing international large-scale assessment data. *Journal of Statistical Software, 81*(7), 1–44.
Checchi, D., & Flabbi, L. (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., & Grand, E. (2017). On estimating achievement dynamic models from repeated cross-sections. *Sociological Methods and Research, 46*(4), 683–714.
Contini, D., & Scagni, A. (2011). Inequality of opportunity in secondary school enrolment in Italy, Germany and the Netherlands. *Quality and Quantity*, *45*, 441–464.
De Gregorio, J., & Lee, J.-W. (2002). Education and income inequality: New evidence from cross country data. *Review of Income and Wealth, 48*(3), 395–416.
Fryer, R. G., & Levitt, S. D. (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*. UK: Final report for the National Equality Panel.
Green, A., Preston, J., & Janmaat, J. (2006). *Education, equality and social cohesion. A comparative analysis*. New York: Palgrave Macmillan.
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. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. 3, pp. 89–200). Amsterdam: North Holland.
Hanushek, E. A., & Woessmann, L. (2015). *The knowledge capital of nations: Education and the economics of growth, CESifo Book Series*. Cambridge: MIT Press.
Heisig, J. P., Schaeffer, M., & Giesecke, J. (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. (Ed.). (2013). *Determined to succeed? Performance versus choice in educational attainment.* Stanford: Stanford University Press.
Jakubowski, M. (2010) Institutional tracking and achievement growth: Exploring difference-in-differences approach to PIRLS, TIMSS, and PISA data. In J. Dronkers (Ed.), *Quality and inequality of education. Cross-national perspectives* (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., Lopez-Agudo, L. A., Marcenaro-Gutierrez, O. D., & Shure, N. (2017). What happens when econometrics and psychometrics collide? An example using the PISA data. *Economics of Education Review, 61,* 51–58.
Jerrim, J., & Micklewright, J. (2014). Socio-economic gradients in children's cognitive skills: Are cross-country comparisons robust to who reports family background? *European Sociological Review, 30*(6), 766–781.
Kerr, S. P., Pekkarinen, T., & Uusitalo, R. (2013). School tracking and development of cognitive skills. *Journal of Labor Economics, 31*(3), 577–602.
Lippman, H., Anderson Moore, K., & McIntosh, H. (2011). Positive indicators of child well-being: A conceptual framework, measures, and methodological issues. *Applied Research Quality Life, 6,* 425–449.
Malamud, O., & Pop-Eleches, C. (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. (2014). *PISA 2012 results in focus. What 15-year-olds know and what they can do with what they know*. https://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 G. J. Duncan & R. J. Murnane (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances.* Russel Sage Foundation.
Ruhose, J., & Schwerdt, G. (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.
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 ed.). Cambridge MA: MIT Press.
Zuckerman, G. A., Kovaleva, G. S., & Kuznetsova, M. I. (2013). Between PIRLS and PISA: The advancement of reading literacy in a 10–15-year-old cohort. *Learning and Individual Differences*, *26*, 64–73.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Submit your manuscript to a SpringerOpen journal and benefit from:**

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

**Submit your next manuscript at ► springeropen.com**