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

Dipartimento
Economia e Statistica
Cognetti de Martiis

Working Paper Series

31/14

CROSS-SECTIONAL LEARNING ASSESSMENTS: COMPARABILITY OF REGRESSION COEFFICIENTS AND VALIDITY OF DIFFERENCE-IN-DIFFERENCE ESTIMATION TO EVALUATE INSTITUTIONAL EFFECTS

DALIT CONTINI

UNIVERSITÀ DEGLI STUDI DI TORINO

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

Cross-sectional learning assessments: comparability of regression coefficients and validity of difference-in-difference estimation to evaluate institutional effects

Dalit Contini

Department of Economics and Statistics "Cognetti de Martiis"- University of Torino

Abstract.

The way achievement inequalities among socio-demographic groups develop throughout childhood in different institutional contexts is a matter of great interest from a policy perspective. Yet, international learning assessments, like many national studies, are cross-sectional and non-vertically-equated (i.e. achievement is not measured on a unique scale in different surveys at different age or grades). Against this background, the aim of this paper is twofold. First, I show that the comparison of regression coefficients with non-vertically-equated achievement scores as dependent variables does not convey much information on the development of disparities related to specific individual characteristic, even when applied to standardized scores. On these grounds, I question the validity of difference-in-difference estimation – whose fundamental element is the comparison of regression coefficients between two non-equated learning assessments administered at different grades – for the evaluation of the impact of institutional features on achievement inequalities related to ascribed individual characteristics like gender or socioeconomic background.

Keywords. Educational economics**,** achievement inequalities, international assessments, cross-sectional data, non-equated scores, difference-in-difference.

JEL classification: I24, C10

1 Introduction

The persistency of educational inequalities among different socioeconomic and demographic groups is an issue of major concern among social scientists. Along large differentials in educational attainment, the development of standardized learning assessments has highlighted the existence of substantial achievement inequalities among children of the same age or school grade in many countries. Moreover, international surveys like PISA, TIMSS and PIRLS have revealed remarkable cross-country variability in the extent to which ascribed individual characteristics affect learning (OECD, 2010a: OECD 2010b; Mullis *et al.* 2012; Mullis *et al.* 2012), and have provided the opportunity to analyze the role played by features of the educational systems in shaping these

inequalities (e.g. Hanushek and Woessmann, 2006; Ammermueller, 2007; Fuchs and Woessmann, 2007; Schuetz *et al.*, 2008).

Given the cross-sectional character of international assessments and of many national studies, inequalities are usually investigated at specific grades or children's age. However, since learning processes are cumulative (Cunha *et al*. 2006) the way inequalities evolve throughout childhood in different institutional contexts is also of great interest. To study the development of inequalities in the absence of longitudinal data, one could compare achievement differentials (or regression coefficients of the relevant socio-demographic variables) over assessments held at different school years. Yet, while some surveys use a single scale to measure achievement at different stages of the schooling career – producing "vertically equated" scores[1] – this is not the case for current international surveys (and many national assessments). Achievement growth cannot be measured with non-equated scores, and direct comparisons across surveys are difficult. To overcome this limit, it is common practice to standardize the scores and compare the corresponding regression coefficients. If the parameter of interest gets larger (smaller), the general interpretation is that there is evidence of increasing (decreasing) inequalities. In this research note, I argue that this strategy may be misleading, as it does not allow isolating the relevant sources of changes – driven by processes that directly involve the socio-demographic characteristic of interest – from other mechanisms affecting the achievement variability over time.

Since achievement inequalities may develop in substantially different ways across educational systems, the economics and sociology of education also deal with the effects of educational policies and institutions on children's learning. In particular, there is a large debate on whether "early tracking" (the differentiation of educational programs at an early age) is detrimental to family background educational inequalities. By exploiting the institutional variability existing at the cross-

[1] For example, vertically equated test scores are provided in the Early Childhood Longitudinal Study-Kindergarten Class of 1998–99 (ECLS–K), a US study focusing on children's early school experiences beginning with kindergarten through middle school. Another example is the Stanford Achievement Test, a standardized achievement tests used by school districts in the United States for assessing children from kindergarten through high school. The primary purpose for creating a single scale for administrative purposes, is to permit users such as school districts with a better means of tracking achievement growth across years and grades (Bielinski *et al*. 2000).

national level, a number of scholars use international assessments to evaluate inequalities across educational systems; in the absence of international longitudinal data, most of these studies focus on inequalities at a given age or school year.

In the attempt to address the problem of confounding effects of other country-specific factors on learning inequalities and improve the estimation of the impact of specific institutions, some papers use difference-in-difference strategies, relating two cross-sectional surveys held at different stages of the schooling career. In their seminal work, Hanushek and Woessmann (2006) apply difference-in-difference to scores' *dispersion* to evaluate whether achievement inequalities in general increase in systems with early tracking relative to comprehensive systems. Drawing on this work, a few scholars – Waldinger (2007), Jakuboski (2010), Van de Werfhost (2013) and Ammermueller (2012) – focus on family background and apply difference-in-difference to the *socioeconomic background regression coefficient*. Despite their limited number, some of these studies are widely cited in the literature, including in the influential Handbook of the Economics of Education (2011)[2], and represent a reference point to researchers aiming at analyzing the effects of institutions on educational inequalities.

In this paper, I address the issue of the validity of the difference-in-difference strategy on *regression coefficients* from an unusual perspective. I do not focus on the identification of the impact of educational policies on inequalities from the point of view of causal inference, where the issues are: Is the causal effect correctly identified? Under what conditions? Instead, I question the validity of the fundamental element of the strategy – the comparison of regression coefficients across surveys – as the basic tool to assess whether specific inequalities increase or not as children age. The central issue here is *how* the development of inequalities are measured, rather than *why* they develop as they do.

[2] Strategy and findings of previous versions of Ammermueller (2012) and Waldinger (2007) are described in the chapters: Hanushek and Woessmann "The economics of international differences in educational achievement" (pg. 156), and Betts "The economics of tracking in education" (pg. 367).

As I show in the first part of the paper, the comparison of regression coefficients on non-equated scores does not convey much information on whether disparities related to particular socio-demographic characteristics have increased or decreased over time. Against this background, the second aim of this note is to question the validity of difference-in-difference techniques applied to *regression coefficients* on dependent variables (typically, international achievement tests) based on different measurement scales. I will show that this approach is clearly incorrect if international scores are used as they are released, i.e. standardized at the cross-country level (as all the papers mentioned above do). Although providing somewhat more meaningful results, difference-in-difference on within-country standardized scores would not be a fully satisfying alternative, since it would not allow isolating the effect of processes involving the specific groups of interest from other mechanisms affecting the scores' variability.

2. A simple student achievement growth model

Consider a stylized model of learning development according to which abilities cumulate over time, so that ability at time *t* equals ability at time *t*-1 plus a growth component.[3] Initial ability and growth may also be affected by individual ascribed characteristics such as gender and family background (e.g. socioeconomic status, minority, ethnic or immigrant origin).

Children from advantaged backgrounds tend to perform better because they live in more stimulating environments and receive more parental support, or because, due to information asymmetries, they are more capable to acquire relevant information on the schooling system and choose better schools. As for gender, the empirical literature shows that girls tend to have lower scores than males on math and higher scores on language. The reasons advanced in the literature refer on the one side to biological and genetic factors, on the other side to the incentives structure (the rational response to the perceived lack of opportunities for women in fields where mathematics

[3] Student Growth Models refer to models of education accountability that measure progress by tracking the achievement scores of the same student from one year to the next with the intent of determining whether the student made progress (Auty 2008). A broad analytical framework for student growth models is that of multilevel modelling (Singer and Willett 2003).

achievement is valued), or to beliefs and stereotypes about status characteristics that may have effects on cognitive performance and self-perception of performance (Penner 2008).

To fix ideas, assume we are interested in gender inequalities.

Suppose we have two cross sectional surveys assessing students' learning at different stages of the educational career, *t*=1 and *t*=2. In order to keep the formalization as simple as possible, I assume that there is no measurement error, so that performance scores can be considered perfect measures of ability. We distinguish the case where the same metric is used to measure achievement at different age or grades, i.e. performance scores are vertically equated,[4] and the case where they are not. Let y_2 be the score at *t*=2 and $\tilde{y}_1$the corresponding vertically equated score at *t*=1. To simplify the exposition, I refer to a single explanatory variable *x* (gender, in our current example) and assume that:

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

and that scores at different ages follow the relation:

$$y_{i2} = \tilde{y}_{i1} + \delta_i$$

where δ_i is achievement growth. If growth is individual-specific and depends linearly on explanatory variables, $\delta_i = \Delta + \beta x_i + \varepsilon_{i2}$. Growth may also depend on previous achievement, so that $\delta_i = \Delta + \beta x_i + \theta \tilde{y}_{i1} + \varepsilon_{i2}$.

However, all international surveys and many national surveys do not equate scales. In this circumstance, $\tilde{y}_1$ represents the (unknown) score at *t*=1 in the measurement scale employed at *t*=2. Assume a linear relation between scales, $\tilde{y}_{i1} = \varphi + \omega y_{i1}$, where y_{i1} is the observed score. Note that φ and ω are not known and not identifiable. The estimable model for y_{i1} is then:

$$y_{i1} = \frac{\tilde{y}_{i1} - \varphi}{\omega} = \frac{\mu_1 - \varphi}{\omega} + \frac{\rho}{\omega} x_i + \frac{\varepsilon_{i1}}{\omega} \quad (2)$$

[4] To create a vertical scale, scores from two tests are linked statistically through a process known as calibration, so that scores can be expressed on a common scale (Patz 2007). TIMSS provides horizontally equated scores (scores of surveys of a given grade at different occasions are equated), but does not provide vertically equated scores (scores of assessments at 4th and 8th grades are not equated).

while the model for y_{i2} is:

$$y_{i2} = \tilde{y}_{i1} + \delta_i = (\varphi + \omega y_{i1}) + \delta_i = (\varphi + \omega y_{i1}) + \Delta + \beta x_i + \theta(\varphi + \omega y_{i1}) + \varepsilon_{i2}$$

$$= \varphi(1 + \theta) + \Delta + \omega(1 + \theta) y_{i1} + \beta x_i + \varepsilon_{i2} \quad (3)$$

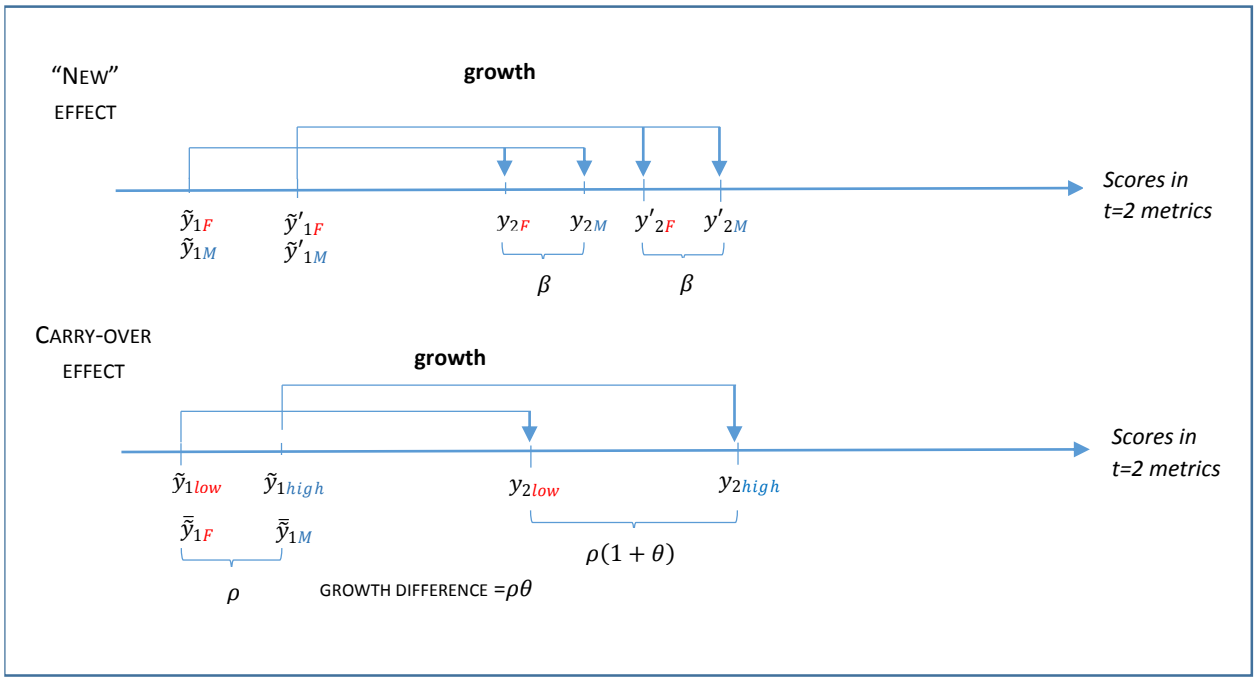
The resulting equation has the structure of a panel data model with a lagged term $y_{i2} = \mu_2 + \gamma y_{i1} + \beta x_i + \varepsilon_{i2}$ with $\gamma = \omega(1 + \theta)$. Notice that γ does not describe the dynamics of the learning process, as it depends on a rescaling factor that allows to translate scores at *t*=1 into scores at *t*=2. Moreover, since θ is unidentified, without vertically equated scores we cannot measure absolute growth, nor test whether achievement of well performing children grows more (or less) than that of lower performing ones.

Substituting (2) into (3) we obtain the cross-sectional model:

$$y_{i2} = \mu_2 + (\beta + (1 + \theta)\rho) x_i + (1 + \theta)\varepsilon_{i1} + \varepsilon_{i2} \quad (4)$$

The cross-sectional regression coefficient $\beta + (1 + \theta)\rho$ represents the overall gender differential developed up to *t*=2. β measures the extent to which the average achievement growth of children with the same score at *t*=1 varies between boys and girls. We will refer to these effects as "new" *x*-effects. These are the most interesting mechanisms because they directly involve gender. Instead, $(1 + \theta)\rho$ are carry-over effects of pre-existing inequalities (see Figure 2).

Figure 2. The two mechanisms underlying the development of *x*-inequalities



3. Assessing the development of inequalities over the child's schooling career

Ideally we would like to: (i) estimate the average achievement growth differential $E(\delta|x+1) - E(\delta|x)$, in our example $E(\delta_M) - E(\delta_F)$, which is equal to $\beta + \theta\rho$; (ii) disentangle the two components, as they capture the effects of substantially different mechanism potentially at work. Under what conditions can we accomplish this goal? It is obviously possible with panel data and equated scores. Instead, with panel data and non-equated scores, we can condition on previous achievement and estimate model (3), and thus identify β. However, we cannot evaluate the average growth differential, because with unknown ω, θ is not identified.

In the following, I analyze whether the development of inequalities can be investigated with *cross-sectional* data by comparing differentials – or regression coefficients – at different occasions. I will show that with equated scores we can evaluate the average differential growth between groups, but not disentangle the two components, while only limited information can be drawn with non-equated scores, even when they are standardized. In particular, none of the relevant mechanism ("new" effects, carryover effects, nor simply the overall effect, i.e. the sum of the two) can be

identified. This discussion will lead us to question the validity of regression coefficients comparison across surveys with non-equated scores as a means to assess how inequalities develop as children age. Identifiable quantities with different data-types and scores are summarized in Table 1.[5]

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

DATA TYPE	SCORES	IND GROWTH δ_i	OVERALL $E(\delta_M) - E(\delta_F)$	"NEW" β	CARRY-OVER $\theta\rho$
Panel data	Equated	YES	YES	YES	YES
	Not equated	NO	NO	YES	NO
Cross-sectional data (comparison of regression coefficients)	Equated	NO	YES	NO	NO
	Not equated-absolute	NO	NO	NO	NO
	Not equated-standardized	NO	NO	NO*	NO

* Although some information may be inferred in some cases (a positive standardized within-country-diff implies $\beta > 0$). See Appendix A.

3.1 Absolute scores

Sticking to the gender inequality example, we may be tempted to evaluate whether in a given country gender inequalities have widened between two assessments, by comparing the average "growth" as measured from observed scores:

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

where x+1 represent males and x females. This amounts to evaluating the difference of regression coefficients at the two assessments, which, from (2) and (4) is: $\beta + (1+\theta)\rho - \frac{\rho}{\omega}$. With vertically equated scores ($\omega = 1$) and growth independent of previous achievement ($\theta = 0$), this quantity – which I name "within-country-diff" – identifies β.

Let us re-write the above expression as $\beta + \theta\rho + (\omega - 1)\frac{\rho}{\omega}$. While β measures whether girls improve or worsen their performance relative to equally performing boys at t=1 (gender-specific mechanisms), $\theta\rho$ are carry-over effects of the preexisting achievement gap (ability-related mechanisms). The term $\beta + \theta\rho$ is the overall difference in the true achievement growth of boys and girls: $E(\delta_M) - E(\delta_F) = (E[y_2|x+1] - E[\tilde{y}_1|x+1]) - (E[y_2|x] - E[\tilde{y}_1|x])$. Thus, with equated

[5] In this paper we do not consider more sophisticated pseudo-panel estimation strategies, that under some conditions allow to estimate β with cross-sectional data (De Simone, 2013, Contini and Grand, 2014).

scores, the within-country-diff captures the overall growth differential. On the contrary, $(\omega - 1)\ \frac{\rho}{\omega}$ has no substantive meaning. Hence, with non-equated scores the within-country-diff is meaningless.

3.2 Standardized scores

The prevalent 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. Goodman *et al*., 2009; Jerrim and Choi, 2013). In a regression framework, this amounts to comparing *x*-coefficients from regressions on standardized scores. Results are often illustrated by simple graphs, and widening *z*-scores differentials across children's characteristics are interpreted as evidence of increasing inequalities.[6]

Indeed, differentials on standardized scores are invariant to the metric of scores at *t*=1. However, the sources of change remain unclear. From (2) and (4) we obtain that the *z*-score differentials:

$$E(z_1|x+1) - E(z_1|x) = \frac{\left(\frac{\rho}{\omega}\right)}{\sigma_{y_1}} = \frac{\rho}{\sigma_{\tilde{y}_1}} = \frac{\rho}{\sqrt{\rho^2\sigma_x^2 + \sigma_{\varepsilon_1}^2}}$$

$$E(z_2|x+1) - E(z_2|x) = \frac{(1+\theta)\rho+\beta}{\sigma_{y_2}} = \frac{(1+\theta)\rho+\beta}{\sqrt{((1+\theta)\rho+\beta)^2\sigma_x^2+(1+\theta)^2\sigma_{\varepsilon_1}^2+\sigma_{\varepsilon_2}^2}} \quad (5)$$

Hence the standardized within-country-diff is:

$$[E(z_2|x+1) - E(z_2|x)] - [E(z_1|x+1) - E(z_1|x)] = \frac{(1+\theta)\rho+\beta}{\sigma_{y_2}} - \frac{\rho}{\sigma_{\tilde{y}_1}} \quad (6)$$

This difference informs on the evolution of the overall *x*-effect, only in the special case $\sigma_{\tilde{y}_1} = \sigma_{y_2}$. Otherwise, its meaning is unclear. As an example, consider the situation where $\beta = 0$ and $\theta = 0$. The above difference becomes:

$$\frac{\rho}{\sqrt{\rho^2\sigma_x^2+\sigma_{\varepsilon_1}^2+\sigma_{\varepsilon_2}^2}} - \frac{\rho}{\sqrt{\rho^2\sigma_x^2+\sigma_{\varepsilon_1}^2}} < 0 \quad (7)$$

In this case the distance between children with different *x* narrows simply because at *t*=2 there is higher variability. No genuine mechanism making a group catching up the previous disadvantage has occurred.

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

What substantive mechanisms may make the score variance increase? Again, assume we are interested in gender inequalities. To keep the notation simple, scores are depicted as being dependent on a single explanatory variable; however, other observed or unobserved factors may be involved. For example, the socioeconomic background (SES), which, incidentally, is likely to be independent of gender, may also play a role. Assume that due to the increasing differentiation of educational programs occurring as children grow up, SES-inequalities widen between the two assessments, in the sense that at t=2 higher SES children will perform better on average than equally well performing children of low SES at t=1. Although this mechanism should *not* affect the average gender growth differential in any way (neither directly nor via previous performance), it will increase the scores' variability at t=2, so that – in relative terms – girls and boys eventually get closer. As a consequence, even if in absolute terms the gender differential does not change, it may decrease relative to the scores' standard deviation.

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

In Appendix B (table A1, case 3), I report the results of a simulation study where I show that if the score variance increases enough between the two assessments, the standardized within-country-diff may be negative even if $\beta > 0$ and $\theta > 0$. Note however that a positive standardized within-country-diff does imply a positive β (see proof in Appendix A).

4. International assessments and the evaluation of institutional effects

International assessments are sometimes employed for national analyses, to evaluate inequalities, school effects, the effects of resources (e.g. class size, teacher pupil-ratio) or schooling policies or

to compare the outcomes in few countries. Yet, providing comparable measures of competencies across countries and covering different schooling systems, international assessments are also increasingly employed to analyze the effects of system-level features of educational systems (for an extensive review, see Hanushek and Woessmann, 2011).

The most common modelling strategies are pooled individual-level models describing performance scores across countries – where institutions are included as country-level explanatory variables – or two-step models – where the parameter of interest is estimated for each country in the first step, and its relation with system-level features is analysed in the second. These analyses focus on the effects of institutions on mean performance (for example Woessmann 2005, Fuchs ad Woessmann, 2007, Woessmann 2010) or on equality of opportunity, usually operationalized as the socio-economic background gradient (Woessmann 2010). Early tracking is the institution that has received the greatest attention. Schuetz *et al.* (2008) with TIMSS and Brunello and Checchi (2007) with IALS, examine single international assessments with a cross-sectional model of achievement in which individual-level family background is interacted with the country-level variable indexing early tracking. This amounts to comparing the family background effect between tracked and non-tracked countries.[7]

In their seminal paper, Hanushek and Woessman (2006) estimate the causal impact of early tracking on reading achievement test scores' dispersion with a simple difference-in-difference model, exploiting two surveys held at different stages of the schooling career: PIRLS (4^{th} grade) and PISA (age 15). The idea is that all children are taught in uniform school type up to fourth grade, while at age 15 only the students in early tracking countries have experienced educational tracking. The essence of their empirical strategy is to compare the change in scores' dispersion that occurs in this period in countries with and without tracking. Since in early tracking countries dispersion increases over time relative to late tracking countries, the conclusion is that early tracking increases inequality.

[7] While Schuetz et al. (2008) find a substantive negative effect of tracking on children's performance at grade eight, Brunello and Checchi (2007) find the opposite effect on adult's cognitive skills.

Going back to the estimation of the effects of institutions on family background inequalities, the use of cross-sectional methods has been criticized because they do not allow controlling for cross-country cultural and societal differences affecting inequalities. Waldinger (2007), for example, argues that parental background effects are larger in early tracking countries even before tracking is enforced, so the effect of tracking cannot be isolated from the effects of unobserved societal differences already in place before tracking. In this perspective, drawing on the work of Hanushek and Woessman (2006), Waldinger (2007) and some other scholars (Jakubowski 2010, Ammermuller, 2012; van de Werfhorst 2013) exploit the data of two student learning assessments held at different children's age, and employ difference-in-difference strategies on family background regression's coefficients.[8]

These scholars' main concern is the identifiability of the institutional effect from the perspective of causal inference. In this respect, Waldinger (2007, pg. 9) writes: "I compare the change between the early and the late test in the importance of family background in early versus late tracking countries. This is a legitimate strategy to control for unobserved country level variables under the identifying assumption that the unobserved country characteristics do not change between the primary and secondary school grades." Similarly, Ammermueller (2012, pg. 192) argues: "The identification strategy utilizes the difference in the dependence between social status and educational outcomes across grades between countries whose institutions have changed between grades and those with no institutional changes across grades. Thereby, country-specific factors besides the schooling system can be largely controlled for, assuming they are identical for students of age nine/ten and fifteen." Likewise, Van de Werfhorst (2013, pg.1) writes: "Difference-in-difference designs are powerful tools to assess effects of institutions, as variation in inequalities at the first moment of observation are considered as given (which result from unobserved factors),

[8] While Ammermueller (2012) and Van de Werfhorst (2013) find that tracking has negative consequence on equality of opportunity of children of different backgrounds, confirming the results of Hanushek and Woessman (2006) on overall dispersion, Waldinger (2007) and Jakubowski (2010) report no strong evidence of tracking effects. The reason of these contrasting findings may lay in the differing surveys employed, set of countries and model specification. For instance, Ammermueller's specification is much more flexible that that adopted by the other scholars. While country-specific fixed effects are included in all models, he also allows for country-specific regression coefficients.

whereas the effect of tracking is examined by the change in inequalities between the first and the second moment of observation". Still, Ammermueller (2012, pg. 191) argues that despite the modeling strategy pays special attention to identification issues and robustness of the results "[…] the results should rather be interpreted as conditional correlations than as causal effects".

These studies also address comparability issues across surveys. Jakubowski (2010), Van de Werfhorst (2013) and Ammermueller (2012) point out that the sample of countries differs in PIRLS, TIMSS and PISA. Thus, they run difference-in-difference regression models on rescaled scores – with same mean and standard deviation among analyzed countries – as there is consensus that this standardization solves comparability problems concerning the different measurement scale across surveys.

Summing up, the limits of difference–in-difference designs to evaluate institutional effects highlighted in the literature refer to the adequacy of causally relating the observed changes in the family background coefficients to schooling system features. However, no discussion is made on the validity of difference-in-difference – based in essence on the comparison of regression coefficients across surveys and countries – to evaluate whether these coefficients actually *measure* what they are supposed to, i.e. whether achievement inequalities between children of differ backgrounds are widening over the child's life course or not.

In the previous section, I have shown that comparing regression coefficients from different assessments is generally not informative on the existence of processes making the achievement of children with certain characteristics grow more or less than that of children with different characteristics. This applies to absolute scores, but also to standardized scores. If the comparison of regression coefficients does not convey much information on the development of *specific* (gender, socioeconomic, ethnic..) inequalities over time, there are good reasons to cast doubts on difference-in-difference strategies, based on the comparison of these coefficients across countries. In the following section, I discuss this issue in analytical terms, while in Appendix C I show the results of

a simulation exercise that allows to better focus on specific examples and that proves the general inadequacy of the approach.

5. Difference-in-difference with international scores

5.1 Performance scores in international assessments

The most clear-cut example of what "absolute scores" are is the percentage of correct answers in a test, or some weighted average of correct and incorrect answers. However, this is a rather outdated way of measuring individual ability. Like many national assessments, all international survey on children's learning rely on Rasch models or Item Response Theory (IRT) to produce measures of achievement. These methods take into account the items' difficulty, and in some cases the guessing probability and the items' discriminatory power.[9] 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.[10] Transformed scores have mean 500 and SD 100.

In this sense, PISA, PIRLS and TIMSS produce standardized scores. Yet, a fundamental difference with the notion of standardized scores employed in section 3 is that we were considering scores that had been standardized within countries. Instead, international assessments use the same yardstick for all countries: if we compare two French individuals in TIMSS, we observe how many SD they are apart with respect to the cross-country SD, not to the French SD. As we will see below, to our end this feature makes international scores more similar to absolute rather than to standardized scores.

5.2 Difference-in-difference on original international scores

What are the implications of the results derived in section 3 on the validity of difference-in-difference strategies to evaluate the effects of early tracking on family background inequalities?

[9] 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.

[10] To be more precise, five random draws (the so-called "plausible values") from the posterior distribution of ability given the item's response pattern are taken for each individual.

To fix ideas, let us think of TIMSS 4th grade as the test at *t*=1 and TIMSS 8th grade as the test at *t*=2. Assume we have only two ideal-type countries, both with comprehensive systems in 4th grade while one with (T) and the other without tracking ($\bar{T}$) in 8th grade. As recalled above, international scores are standardized so to have overall mean 500 and average country SD 100.[11] Measurement scales are non-equated. Thus, we may think of the original (equated) ability measures a_{it} of individuals *i* at time *t* as being generated from models (1) and (3) – with *a*s substituting *y*s. Roughly speaking, these ability measures are translated into international scores according to: $y_{it} = \left(\frac{a_{it}-\bar{a}_t}{\sigma_{a_t}(cc)}\right)100 + 500$, where $\sigma_{a_t}(cc)$ is the average cross-country SD of the original measure. In the subsequent, I will refer to these as "TIMSS-like scores".

Consider a simple case where model parameters capturing family background and prior achievement effects vary across countries only depending on whether there is early tracking or not, while the others (intercepts and variance of the error term) are country specific. To make this clear, we denote with the subscript *track* the corresponding coefficients. For each country *c*, the TIMSS-like scores' model at *t*=1 is

$$y_{i1c} = \mu'_{1c} + \rho_{track}F_1x_{ic} + \varepsilon'_{i1c} \quad (8)$$

while at *t*=2 is

$$y_{i2c} = \Delta'_c + (\beta_{track} + (1 + \theta_{track})\rho_{track})F_2x_{ic} + \varepsilon'_{i2c} \quad (9)$$

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

It should be clear that despite standardization, international scores are not "standardized" in the sense of section 3. Each assessment has its own metric (estimable coefficients depend on the scale *F*), thereby, as discussed in section 3.1, the comparison of regression coefficients is non-informative on the development of inequalities.

In essence, what difference-in-difference strategies do, is to evaluate:

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

$$DID = [(\beta_T + (1+\theta_T)\rho_T) - (\beta_{\bar{T}} + (1+\theta_{\bar{T}})\rho_{\bar{T}})]F_2 - (\rho_T - \rho_{\bar{T}})F_1 \quad (10)$$

The building blocks of *DID* are regression coefficients expressed in each assessments' metrics, according to (8) and (9). It is clear that *DID* does not allow to identify any of the relevant mechanisms at play, because it depends on the relative variability of ability at the two assessments.

Still, one might be interested in investigating whether the combined empirical evidence on all regression coefficients and *DID* allows to say something meaningful on the sign of the relevant parameters β and θ. The general answer is no. The relevant case-types are described in Table 2. While in most cases the observed results imply that either β or θ must be positive (or negative), nothing can be said *a priori* when the regression coefficients differentials have the same sign and *DID* has the opposite sign.[12]

To examine this matter in more detail, I developed a simple simulation exercise, summarized in Table 3, focusing on situations 3 and 4. In all cases, the difference between regression coefficients of tracked and non-tracked countries is non-negative at both assessments. In the upper panel I show the model parameters and original scores' mean and SD. In the lower panel, I report descriptive statistics on TIMSS-like transformed scores, separately for tracked and non-tracked countries, and the relevant empirical evidence: the difference between the *x*-regression coefficients (RCΔ1 and RCΔ2) of cross-sectional models (tracked minus non-tracked) and the corresponding *DID*.

Table 2. Empirical evidence and implications on β and θ

SITUATION	Regression coefficients at *t*=1	Regression coefficients at *t*=2	DID	IMPLICATIONS
1	$\rho_T < \rho_{\bar{T}}$	$\beta_T + (1+\theta_T)\rho_T = \beta_{\bar{T}} + (1+\theta_{\bar{T}})\rho_{\bar{T}}$	Necessarily positive	$(\theta_T > \theta_{\bar{T}}) \cup (\beta_T > \beta_{\bar{T}})$
2	$\rho_T = \rho_{\bar{T}}$	$\beta_T + (1+\theta_T)\rho_T > \beta_{\bar{T}} + (1+\theta_{\bar{T}})\rho_{\bar{T}}$	Necessarily positive	$(\theta_T > \theta_{\bar{T}}) \cup (\beta_T > \beta_{\bar{T}})$
3	$\rho_T > \rho_{\bar{T}}$	$\beta_T + (1+\theta_T)\rho_T > \beta_{\bar{T}} + (1+\theta_{\bar{T}})\rho_{\bar{T}}$	Positive	$(\theta_T > \theta_{\bar{T}}) \cup (\beta_T > \beta_{\bar{T}})$
4	$\rho_T > \rho_{\bar{T}}$	$\beta_T + (1+\theta_T)\rho_T > \beta_{\bar{T}} + (1+\theta_{\bar{T}})\rho_{\bar{T}}$	Negative	none
5	$\rho_T < \rho_{\bar{T}}$	$\beta_T + (1+\theta_T)\rho_T > \beta_{\bar{T}} + (1+\theta_{\bar{T}})\rho_{\bar{T}}$	Necessarily positive	$(\theta_T > \theta_{\bar{T}}) \cup (\beta_T > \beta_{\bar{T}})$

[12] The proof is trivial and not reported here. It is available from the author upon request.

Table 3. Difference-in-difference on TIMSS-like scores

Case	Original model parameters Tracked country					Original model parameters Non-Tracked country					Original scores cross-country mean		Original scores cross-country SD	
	ρ	β	θ	$\sigma_{\varepsilon 1}$	$\sigma_{\varepsilon 2}$	ρ	β	θ	$\sigma_{\varepsilon 1}$	$\sigma_{\varepsilon 2}$	t=1	t=2	t=1 $\sigma_{\tilde{y}1}(cc)$	t=2 $\sigma_{y2}(cc)$
1	30	-10	0	40	20	20	0	0	40	20	312	510	42.0	45.8
2	30	0	0	20	40	20	0	0	20	40	312	512	23.7	46.5
3	30	10	0	20	100	20	0	0	20	100	312	515	23.7	103.3
4	30	10	0	20	20	20	0	0	20	20	312	515	23.7	32.3
5	30	0	0.4	20	20	20	0	0	20	20	312	525	23.7	35.1
6	30	30	0	20	90	20	0	0	20	20	312	520	23.7	63.4

ω=0.5, σ_x^2=0.25, $\mu_1 = 300$, $\Delta = 200$ (exception: case 5 tracked $\Delta = 100$)

Case	TIMSS-like scores within country mean				TIMSS-like scores within country SD				Empirical evidence on TIMSS-like scores		
	t=1 T	t=2 T	t=1 $\bar{T}$	t=2 $\bar{T}$	t=1 T	t=2 T	t=1 $\bar{T}$	t=2 $\bar{T}$	RCΔ1	RCΔ2	DID (1-2)
1	506	500	494	500	102	100	98	100	23.8	0	-23.8
2	510	505	489	495	106	101	94	98	42.2	21.5	-20.7
3	510	505	489	495	106	101	94	99	42.2	19.4	-22.8
4	510	515	489	484	106	107	94	93	42.2	61.9	19.7
5	510	544	489	456	106	115	94	85	42.2	62.7	20.5
6	510	516	489	484	106	153	94	47	42.2	63.1	20.9

*Means and SDs results from a simulation with n=1000000 per country.
Empirical evidence results computed analytically.

** RCΔ=regression coefficient difference
RCΔ1: $(\rho_T - \rho_{\bar{T}})\frac{100}{\sigma_{\tilde{y}1}(cc)}$; RCΔ2: $[(\rho_T(1+\theta_T)+\beta_T) - (\rho_{\bar{T}}(1+\theta_{\bar{T}})+\beta_{\bar{T}})]\frac{100}{\sigma_{y2}(cc)}$

Comparing specific subgroups of cases allows to highlight the low-meaningful information content provided by *DID*.[13]

Cases 1-3: *Negative DID*

Very different mechanisms underlie this result. In case 1 there are negative "new" effects (reducing inequalities) in tracked countries. In case 2 there are no real effects ("new" or carryover) but higher variability at t=2. In case 3 there are positive "new" effects (widening inequalities) in tracked countries and much higher variability at t=2.

Cases 4-6: *Positive DID*

Very different mechanisms underlie this result. In case 4 there are positive "new" effects and fairly stable variability across assessments. In case 6 there are very large positive "new" effects and much

[13] Note that reported DIDs are of similar magnitude and would be interpreted as increasing or decreasing inequalities by approximately 0.2 SD.

higher variability at *t*=2. In case 5 there are no "new" effects but large carryover effects for tracked countries.

Cases 3 and 4: *same β and θ*

Despite their equivalence (β =10 in tracked countries and 0 otherwise, θ=0), *DID* is positive in case 4 but negative in case 3. The negative value in case 3 occurs because there is much higher variability at *t*=2: despite "new" positive effects occurring in tracked countries and the absence of any *real* effect in non-tracked countries, the *relative x*-gap differential between tracked and non-tracked countries (relative to the cross-country mean SD) decreases between the assessments.

5.3 Difference-in-difference on within-country standardized international scores

In Section 3.2 I argued that comparing *x*-differentials of standardized scores at different occasions gives a descriptive picture of how *x*-differentials evolve in a given country, but does not provide information on the sources of the observed change, which may be entirely due to processes not involving *x* nor any variable correlated to *x*. In this perspective, I now examine whether implementing difference-in-difference on within-countries standardized scores would provide meaningful results.[14]

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

$$DID_{st} = \left(\frac{(\beta_T+(1+\theta_T)\rho_T)F_2}{\sigma_{y_2T}} - \frac{((\beta_{\bar{T}}+(1+\theta_{\bar{T}})\rho_{\bar{T}}))F_2}{\sigma_{y_2\bar{T}}}\right) - \left(\frac{\rho_T F_1}{\sigma_{y_1T}} - \frac{\rho_{\bar{T}}F_1}{\sigma_{y_1\bar{T}}}\right) \quad (11)$$

It is immediately clear that the relevant parameters are not identifiable. Yet, DID_{st} is not a meaningless quantity: by dividing by the country score SD we obtain a metric-free measure, thus all terms of (11) are comparable: in a purely descriptive perspective, some conclusions might still be drawn.

[14] Within-country standardized scores are obtained by multiplying international scores by $100/\sigma_c$, where σ_c is the country-specific SD.

Let us return to the gender inequality example. Consider a negative DID_{st}. This result provides evidence that *for some reason*, the gender relative gap (relative to *each* country's SD) has increased more (or decreased less) in non-tracked countries than in tracked countries. As discussed above, there are various possible reasons underlying this empirical result: different "new" gender effects given previous ability between tracked and non-tracked countries, different carryover effects of prior ability, different effects of other explanatory variables that may influence the within-country scores' variability. For example, if SES inequalities get larger after tracking, other thing being equal, the *relative* gender difference should decrease in tracked countries relative to non-tracked countries.

It is noteworthy to mention that the use of within-country standardized scores may yield to very different DID_{st} as compared to raw international scores. To get an intuition of the way they behave, I compare *DID* and DID_{st} in the same cases of Table 3. Results are shown in Appendix C.

6. Concluding remarks

In this paper I make two main points. (i) I show that the comparison of regression coefficients on non-vertically-equated achievement scores does not convey much information on whether disparities related to a particular socio-demographic characteristic increase or not as children age, the reason being that it does not allow isolating the effect of processes involving the socio-demographic characteristic of interest from other mechanisms affecting the scores' variability. (ii) With reference to difference-in-difference strategies employed to estimate the effect of institutional features on inequalities related to specific socio-demographic characteristics (for example, family background), I question the validity of difference-in-difference techniques applied to regression coefficients when the dependent variables – here, international achievement tests – are not measured on a unique scale. I show that the approach is patently wrong if international scores are used as they are released, and has major shortcomings even if applied to within-country standardized scores. It is worthwhile noticing that this criticism does not apply to difference-in-

difference on dispersion measures as done by Hanushek and Woessmann (2006), because these scholars analyze changes in variability as such, and do not attempt to ascribe the observed changes to the widening or narrowing of differentials involving specific independent variables.

References

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

Ammermueller, A. (2012) Institutional features of schooling systems and educational inequality: cross-country evidence from PIRLS and PISA, *German Economic Review*, **14**(2): 190-213

Auty W. (2008) Implementer's Guide to Growth Models, CCSSO-Council of Chief State School Officers, Washington, DC.

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.

Bielinski J., Thurlow M., Minnema J., Scott J. (2000) How out-of-level testing affects the psychometric quality of test scores. *Out-of-Level Testing Report 2. National Center on Educational Outcomes, University of Minnesota*

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

Contini D., Grand E. (2013). On estimating achievement dynamic models from repeated cross-sections, *Working Paper of the Department of Economics and Statistics Cognetti de Martiis, 43/13*

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

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

Fuchs, T. and 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. and Washbrook, E. (2009) Inequalities in educational outcomes among children aged 3 to 16. *Final report for the National Equality Panel*, UK

Hanushek, E.A. and Woessmann, L. (2006) Does educational tracking affect performance and inequality? Differences-in-differences 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.

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

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

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

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

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

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

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

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

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

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

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

Woessmann L. (2010) Institutional Determinants of School Efficiency and Equity: German States as a Microcosm for OECD Countries, *Jahrbücher für Nationalökonomie und Statistik*, 230(2), 234-270.

Appendix A

Proof that a positive within-country-diff implies a positive β.

A positive difference in regression coefficients (within-country-diff) implies that:

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

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

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

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

Let us consider now models (1) and (4) and derive the corresponding score variances.

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

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

Their ratio is:

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

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

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

Hence, for $\rho > 0$ and under the (highly reasonable) assumption that $\theta > -1$, $\frac{\sigma_{y_2}}{\sigma_{\tilde{y}_1}} > (1+\theta)$.

In conclusion, result (A.1) implies $\beta > 0$.

Appendix B.

Comparing regression coefficients of different assessments in a single country

Table A1. Country-diff on absolute and standardized scores

Case	Model parameters					SD		Regr. coeff. Absolute		Country-diff Abs.	Regr. coeff. Stand.		Country-diff Stand.
	ρ	θ	β	$\sigma_{\varepsilon 1}$	$\sigma_{\varepsilon 2}$	$\tilde{y}_1$	y_2	y_1	y_2	y	z_1	z_2	Z
1	20	0	0	30	10	32	33	40	20	-20	0.63	0.60	-0.03
2	20	0	0	30	30	32	44	40	20	-20	0.63	0.46	-0.17
3	20	0.05	5	30	50	32	61	40	26	-14	0.63	0.43	-0.20
4	20	0.05	5	30	30	32	45	40	26	-14	0.63	0.57	-0.06
5	20	0.05	5	30	10	32	36	40	26	-14	0.63	0.73	0.10
6	20	0.1	10	30	40	32	54	40	32	-8	0.63	0.59	-0.04
7	20	0.1	10	30	34	32	50	40	32	-8	0.63	0.64	0.01
8	20	0.1	10	30	15	32	40	40	32	-8	0.63	0.81	0.18
9	20	-0.05	-5	30	30	32	42	40	14	-26	0.63	0.33	-0.30
10	20	-0.05	-5	30	10	32	31	40	14	-26	0.63	0.45	-0.18
11	20	0.1	0	30	10	32	36	40	22	-18	0.63	0.61	-0.02
12	20	0.5	0	30	10	32	48	40	30	-10	0.63	0.62	-0.01

ω=0.5, σ_x^2=0.25

Appendix C.

Difference-in-difference on within-country standardized scores.

Table A2. *DID* on absolute and within-country standardized scores

Case	Original model parameters Tracked country					Original model parameters Non-Tracked country					TIMSS -like scores	Empirical evidence on within country standardized TIMSS-like scores				
	ρ	β	θ	$\sigma_{\varepsilon 1}$	$\sigma_{\varepsilon 2}$	ρ	β	θ	$\sigma_{\varepsilon 1}$	$\sigma_{\varepsilon 2}$	DID	RC1T	RC1$\bar{T}$	RC2T	RC2$\bar{T}$	DID
1	30	-10	0	40	20	20	0	0	40	20	-23.8	0.70	0.49	0.44	0.44	-0.21
2	30	0	0	20	40	20	0	0	20	40	-20.7	1.20	0.89	0.64	0.44	-0.11
3	30	10	0	20	100	20	0	0	20	100	-22.8	1.20	0.89	0.38	0.20	-0.12
4	30	10	0	20	20	20	0	0	20	20	19.7	1.20	0.89	1.15	0.67	0.18
5	30	0	0.4	20	20	20	0	0	20	20	20.5	1.20	0.89	1.04	0.67	0.07
6	30	30	0	20	90	20	0	0	20	20	20.9	1.20	0.89	0.62	0.67	-0.35

ω=0.5, σ_x^2=0.25, $\mu_1 = 300$, $\Delta = 200$ (exception: case 5 tracked $\Delta = 100$)

*Means and SDs results from a simulation with n=1,000,000 per country.
Empirical evidence results computed analytically.

** Regression Coefficient 1: $\frac{\rho_T}{\sigma_{\tilde{y}1T}}$ or $\frac{\rho_{\bar{T}}}{\sigma_{\tilde{y}1\bar{T}}}$ Regression Coefficient 2: $\frac{\beta_T+(1+\theta_T)\rho_T}{\sigma_{y2T}}$ or $\frac{(\beta_{\bar{T}}+(1+\theta_{\bar{T}})\rho_{\bar{T}})}{\sigma_{y2\bar{T}}}$

Recall that the metric of the TIMSS-like scores *DID* is in actual score points (*DID*=20 means that it is 0.2 times the average cross-country SD), while DID_{st} is already expressed in SD units (the own country's SD). Results on the selected cases show that in most cases DID_{st} has the same sign of the original *DID* (although their size may differs considerably). In case 6, however, the original *DID* is positive whereas the DID_{st} is negative (and large in size). This is because the unexplained variability rises between assessments much more in tracked countries than in non-tracked countries: relative to each country's SD, the same *x*-gap will weight much less in the former.