On estimating achievement dynamic models from repeated cross-sections

Department of Economics and Statistics
"COGNETTI DE MARTIIS"

Working Paper Series

43/13

ON ESTIMATING ACHIEVEMENT DYNAMIC MODELS FROM REPEATED CROSS-SECTIONS

DALIT CONTINI and ELISA GRAND

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

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

On estimating achievement dynamic models from repeated cross-sections

Dalit Contini
University of Torino

Elisa Grand
University of Torino

Summary. Despite the increasing spread of standardized assessments of student learning, longitudinal achievement data are still lacking in many countries. This article raises the following question: can we exploit cross-sectional assessments held at different schooling stages to evaluate how achievement inequalities related to individual ascribed characteristics develop over time? We discuss the issues involved in estimating dynamic models from repeated cross-sectional surveys and, consistently with a simple learning accumulation model, we propose an imputed regression strategy that allows to "link" two surveys and deliver consistent estimates of the parameters of interest. We then apply the model to Italian achievement data of 5th and 6th graders and investigate how inequalities develop between primary and lower secondary school.

Keywords: achievement inequalities, dynamic models, pseudo-panel estimation, repeated cross-sections, standardized assessments.

1. Introduction

The expansion of standardized learning assessments at the national and international level has fostered the study of educational inequalities in terms of achievement and acquired competences. International surveys like PISA, TIMSS and PIRLS[1] have also given the opportunity to highlight remarkable cross-country variability in the extent to which ascribed individual characteristics such as gender and family background affect learning (OECD, 2010a: OECD 2010b; Mullis *et al.* 2012; Mullis *et al.* 2012), and to relate these differences to schooling policies and features of the educational systems (eg. Hanushek and Woessmann, 2006; Ammermueller, 2007; Fuchs and Woessmann, 2007; Schuetz *et al.*, 2008).

International assessments and many national studies, however, are cross-sectional. In this context, inequalities can only be investigated at specific grades or children's age. Yet, as emphasized by Cunha *et al*. (2006), learning processes are cumulative. Thus, greater knowledge of how differentials across socio-demographic groups evolve throughout childhood in different institutional contexts could help the design of effective educational policies to contrast inequalities.

This article raises the following question: in the absence of longitudinal data, can we exploit cross-sectional standardized assessments held at different stages of the schooling career to evaluate how learning inequalities *develop* over children's life course?

Since different assessments are often not directly comparable, the existing literature has addressed this issue by computing standardized scores and comparing the average *z*-score of individuals of different backgrounds as children age (Goodman *et al.*, 2009; Jerrim and Choi, 2013). Widening *z*-scores differentials across socioeconomic backgrounds are interpreted as evidence of increasing inequalities. Yet, this method does not allow to distinguish between the direct effect of socio-demographic variables operating at each stage of schooling and carryover effects of preexisting gaps; it is also influenced by measurement errors.

In the absence of panel data, individuals cannot be traced over time. The econometric literature offers a number of contributions on the estimation of models for panel data from repeated cross-sections (Deaton, 1985; Moffitt, 1993; Verbeek and Vella, 2005); as shown by Verbeek and Vella (2005), the conditions for consistent estimation are unrealistic in many contexts.

Drawing from this body of work, we discuss the issues involved in estimating dynamic models from repeated standardized cross-sectional surveys on educational achievement, with the aim to estimate how inequalities across socio-demographic groups develop over stages of schooling. We

[1] PISA (*Programme for International Student Assessment*) is conducted by OECD. TIMSS (*Trends in Mathematics and Science Study*) and PIRLS (*Progress in International Reading Literacy Study*) are promoted by IEA, the International Association for the Evaluation of Educational Achievement. PIRLS evaluates children of grade 4, TIMSS focuses on grades 4 and 8, PISA on children of age 15, regardless of the grade attended.

argue that the model allowing to address this research question is very simple, and therefore the conditions for consistent estimation are met. Coherently with a basic learning accumulation model, we propose an imputed regression strategy that allows to "link" two assessments held at different grades. In essence, true lagged values are substituted with appropriate estimates derived from the first survey. The main drawback of imputed regression, however, is that due to this substitution, standard errors of the estimates are greatly inflated. Imputed regression has an advantage over genuine panel data models: by explicitly addressing the issue of measurement error, it provides consistent estimates of the parameters of the model of interest even with an additional source of error, i.e. test scores imperfectly measuring achievement.

In the empirical application we exploit the dataset of the Italian learning assessment of reading and math literacy, carried out by the National Evaluation Agency (INVALSI) on 5^{th} and 6^{th} graders in 2010 and 2011 on a sample of more than 30.000 pupils. We investigate gender, socioeconomic, immigrant background and territorial inequalities at the transition between primary and lower secondary school. This is a relevant turning point, as secondary school becomes much more demanding in terms of achievement requirements.

Our contribution to the existing literature is threefold. Firstly, we provide a reflection on the advantages of pseudo-panel modelling for the study of the development of learning inequalities over stages of schooling, and the conditions for consistent estimation. If repeatedly applied to different segments of the schooling career, this technique allows to investigate how inequalities develop over children's educational life course, moving the focus of the literature on achievement inequalities from a static to a dynamic perspective. Secondly, we substantiate our theoretical results with simulations, and show that large samples and good instruments are needed to obtain reliable results. Thirdly, by exploiting a large scale national standardized assessment held at different grades, we analyze how inequalities evolve between primary and lower secondary school in Italy and show that socio-demographic differentials amplify in reading, while the North-South divide severely widens in math.

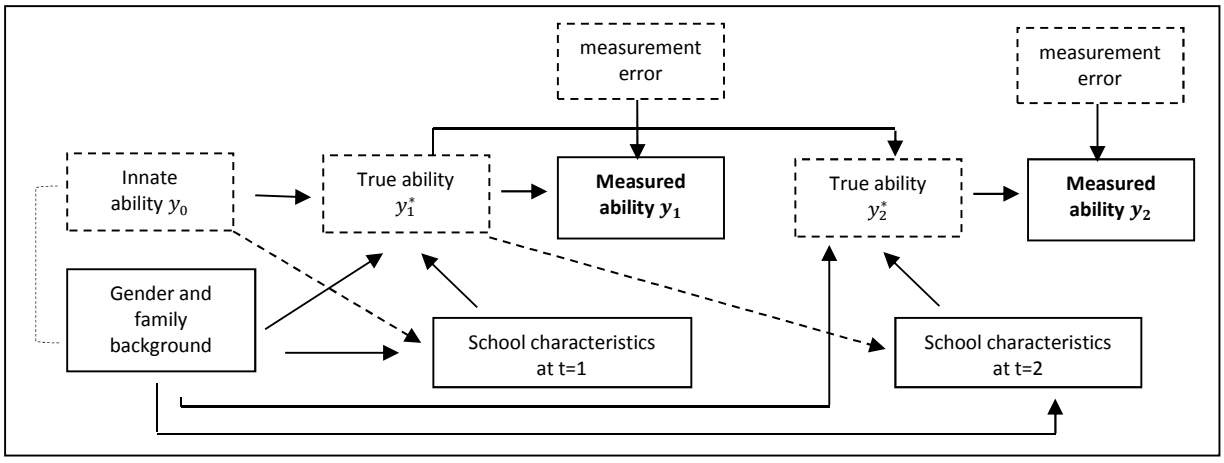
The paper is organized as follows. In section 2 we define the model. In section 3 we elaborate on a fully cross-sectional approach, diff-in-diff estimation, and show that this strategy does not allow to identify the sources of dynamics of achievement inequalities. In section 4 we present and discuss the imputed regression estimation strategy. Our data and case-study are presented in section 5. Conclusions follow. In the Appendix we present a simulation study designed to evaluate the order of magnitude of standard errors of the estimates in the imputed regression strategy, assess the behavior of alternative estimation strategies and evaluate results in the presence of children repeating grades.

2. The model

Consider two cross-sectional surveys assessing students' learning at times *t*=1 and *t*=2. A stylized but comprehensive model of learning development and observed performance scores, consistent with the idea of a cumulative process where abilities build up over time, is depicted in Figure 1. Innate ability could be independent of individual characteristics, but this condition is not necessary. True unmeasured ability follows a Markov process, as ability at time *t* depends on ability at time *t*-1, but not on previous ability. Test scores (measured ability) are additive functions of true ability and an independent measurement error. True ability is affected by individual variables such as gender and family background (say, socioeconomic status and immigrant origin).

Children from advantaged backgrounds perform better on average because they usually live in a more culturally stimulating environment and receive more parental support, but also because they may choose better schools. School choices may also be driven by children's ability. We assume that school characteristics at *t*=1 do not directly affect ability at *t*=2, given ability at *t*=1.

Figure 1. A stylized dynamic model of performance scores



NOTE. Solid boxes represent observable variables, dashed boxes unobservables. Solid arrows represent well-established causal relations. Dashed arrows stand for causal relations which might exist or not (depending for example on the educational system). Curved dashed lines represent possible correlations between variables.

In the above framework, we consider the following underlying linear autoregressive model:

$$y_{it} = \mu_t + \gamma_t y_{it-1} + \beta_t' \boldsymbol{x}_i + \varepsilon_{it} \quad (1)$$

where y_t and y_{t-1} represent performance scores at two moments of the schooling career, and $\boldsymbol{x}$ is a vector of socio-demographic individual variables. School characteristics are not included among the explanatory variables, the reason being that our interest rests on *inequalities*, i.e. on the total effect of socio-demographic variables, given by direct effects and indirect effects through school features. If children from advantaged backgrounds choose better schools, adding school variables would capture part of the desired effect. Similarly, we deliberately exclude other intervening variables

such as intentions, aspirations, learning strategies, behaviors. As a consequence, the set of explanatory variables of interest consists of (nearly) time-invariant socio-demographic characteristics. The error term is uncorrelated over time and individuals, and is independent of $\boldsymbol{x}$ and of the lagged score.

We assume time-varying parameters firstly because assessments administered at different grades are not necessarily separated by a uniform time span; more importantly, because there is no reason to assume that the relation between scores of subsequent assessments and the effect of explanatory variables are constant over the schooling career. Indeed, our aim is to study how inequalities develop over time.

The fact that parameters are allowed to change over time implies that (unless we make strong assumptions on how they evolve) we must consider two assessments at a time, and estimate the model for each pair of subsequent assessments. Considering a scalar explanatory variable to simplify the exposition, with two waves general model (1) reduces to:

$$y_{i1} = \mu_1 + \rho x_i + \varepsilon_{i1} \tag{2}$$

$$y_{i2} = \mu_2 + \gamma y_{i1} + \beta x_i + \varepsilon_{i2} \tag{3}$$

The error terms include a random component with the usual properties and measurement error, assumed to be independent of true scores. ε_1 also captures innate ability.[2] ρ and β are measures of learning inequality. The parameter of main interest is β, representing differentials developing between times *t*=1 and *t*=2, on top of those already in place at *t*=1. If $\rho \neq 0$ and $\beta = 0$ the explanatory variable affects achievement up to *t*=1, but given achievement at *t*=1, on average at *t*=2 children of different backgrounds reach the same performance level. On the other hand, if ρ and β have the same sign inequalities widen; if they have opposite signs, they weaken or change direction.

Notice that we do not assume a conventional static model for panel data with individual fixed effects such as $y_{it} = \mu_t + \alpha_i + \beta_t' \boldsymbol{x}_i + \varepsilon_{it}$, because the autoregressive model is theoretically better suited to represent a cumulative learning process where competencies build up over time. In addition, at *t*=2 this model is a particular case of (3) with $\gamma = 1$. We do not consider a dynamic model with fixed-effects either, firstly because with two points in time the model would be unidentified, secondly because the fixed effect component is redundant if we conceive it as innate ability (the assumed Markov structure posits that ability at *t*=2 does not depend on innate ability given ability at *t*=1).

[2] According to (1), $y_{i,t-1} = \mu_1 + \gamma_{t-1} y_{i,t-2} + \beta_{t-1}' \boldsymbol{x}_i + \varepsilon_{i,t-1}$. Going backwards and making repeated substitutions, y_{t-1} can be expressed as a function of innate ability y_0 that, being unobservable, enters the error term. The resulting equation has the structure of (2).

2.1 An alternative derivation of the model

In the previous section we have specified the panel data model as conventionally done in econometrics. We now derive the model from the perspective of individual growth models.[3] Consider first a set of cross sectional assessments where the metric in which the achievement is measured is preserved across time, i.e. performance scores are "vertically equated".[4] In this case subsequent scores follow the relation: $y_{i2} = y_{i1} + \delta_i$, where δ_i is achievement growth. If growth is individual-specific and depends on explanatory variables, $\delta_i = \Delta + \beta x_i + \varepsilon_{i2}$ and $y_{i2} = y_{i1} + \Delta + \beta x_i + \varepsilon_{i2}$. Yet, growth may also depend on previous achievement. In this case $y_{i2} = y_{i1} + \Delta + \beta x_i + \theta y_{i1} + \varepsilon_{i2}$.

On the contrary, if achievement scores are not equated, the relation between subsequent scores would be: $y_{i2} = \tilde{y}_{i1} + \delta_i$, where $\tilde{y}_{i1} = \varphi + \omega y_{i1}$ represents achievement at *t*=1 in the measurement scale employed at *t*=2. Since φ and ω are not known and not identifiable, we cannot measure absolute growth, but only evaluate individuals' position relative to others. In the general case, the model then becomes:

$$y_{i2} = \varphi(1+\theta) + \Delta + \omega(1+\theta)y_{i1} + \beta x_i + \varepsilon_{i2} \quad (4)$$

The resulting model has the structure of (3), 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 test whether achievement of well performing children grows more (or less) than that of lower performing ones.[5]

3. Cross-sectional strategies: difference in difference

Model (2) can be estimated with conventional methods using the cross-sectional survey at *t*=1.[6] We now derive the conditions for consistent estimation of β in model (3) with a simple fully cross-sectional diff-in-diff strategy, and show that it produces consistent estimates of β only in the

[3] Growth models analyze an outcome variable measured at repeated occasions and model it as a function of time (Singer, Willett, 2003). They are often used in the statistic-educational literature for accountability purposes, to evaluate school effectiveness and assess the impact of specific educational programs.

[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. TIMSS provides horizontally equated scores (scores of surveys of a given grade at different occasions are equated), but does not equate scores of assessments of different grades. The Italian survey employed in our empirical analysis does not equate scores, neither horizontally, nor vertically.

[5] Due to the issue of "regression to the mean", in the presence of measurement error in test scores the effect of previous performance would be difficult to identify even with equated scores (Jerrim and Vignoles, 2013). Ceiling effects may also operate (Betebenner and Linn, 2010).

[6] If innate ability is independent of *x*, $\hat{\rho}$ captures the effects of family background related to environmental and cultural factors. If, as maintained by some scholars, the assumption is not valid, $\hat{\rho}$ will also capture genetic effects. Note that this issue is not relevant for the estimation of (3), as in this case innate ability is entirely captured by y_1.

simplest situation, where scores are vertically equated ($\omega = 1$) and growth does not depend on previous achievement ($\theta = 0$).

Diff-in-diff amounts to computing the difference between regression coefficients of the cross-sectional models relative to the two assessments:

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

It is trivial to show that when $\omega = 1$ and $\theta = 0$ this difference is equal to β. In all other cases diff-in-diff fails to identify β. With vertically equated scores, if growth depends on previous achievement diff-in-diff equals $\beta + \theta\big[E[y_1|x+1] - E[y_1|x]\big]$.

If achievement scores are not vertically equated, in the simplest model with $\theta = 0$ diff-in-diff equals $\beta + (\omega - 1)\,(E[y_1|x+1] - E[y_1|x])$. Not even the sign of β can be predicted, as there is no prior knowledge on ω (which could be smaller or larger than 1). Thus, diff-in-diff may be positive with nil or negative β and vice-versa. In the general case where growth depends on previous scores, diff-in-diff amounts to:

$$\beta + \omega\theta(E[y_1|x+1] - E[y_1|x]) + (\omega - 1)\,(E[y_1|x+1] - E[y_1|x]) \quad (5)$$

The first two terms represent the overall achievement growth differential:[7]

$$(E[y_2|x+1] - E[\tilde{y}_1|x+1]) - (E[y_2|x] - E[\tilde{y}_1|x])$$

β captures the differential developed between the two assessments that can be directly ascribed to X, while $\omega\theta(E[y_1|x+1] - E[y_1|x])$ captures the effects driven by the preexisting achievement gap. The last term of (5), instead, has no substantive meaning. Therefore, diff-in-diff on absolute scores does not convey any useful information on the development of inequalities, as it fails to identify β – which accounts for new inequalities developed between *t*=1 and *t*=2 – and the overall achievement growth differential – which also incorporates carryover effects of preexisting gaps.

Diff-in-diff on standardized scores

To overcome the difficulties in comparing test scores of different assessments, the principal strategy adopted in the existing literature studying the evolution of inequalities over childhood is to standardize scores and compare average *z*-scores of individuals of different backgrounds as children age (Goodman et al., 2009; Jerrim and Choi, 2013). This amounts to applying diff-in-diff on standardized scores.[8] Results are typically illustrated by simple graphs: widening *z*-scores

[7] $(E[y_2|x+1] - E[\tilde{y}_1|x+1]) - (E[y_2|x] - E[\tilde{y}_1|x]) = (E[y_2|x+1] - E[y_2|x]) - (E[\tilde{y}_1|x+1] - E[\tilde{y}_1|x]) = \beta + (1+\theta)(E[\tilde{y}_1|x+1] - E[\tilde{y}_1|x]) - (E[\tilde{y}_1|x+1] - E[\tilde{y}_1|x]) = \beta + \omega\theta(E[y_1|x+1] - E[y_1|x])$.

[8] Note that applying diff-in-diff on international scores is not equivalent to applying diff-in-diff on standardized scores. International assessments PIRLS, PISA and TIMSS do provide standardized scores (with mean 500 and st.dev. 100), but the standardization is performed with reference to a set of countries, varying over time and across surveys.

differentials across socioeconomic backgrounds are interpreted as evidence of increasing inequalities.[9]

Indeed, diff-in-diff on standardized scores is invariant to the metric of scores at *t*=1. However, the sources of change remain unclear. Let us analyze the differentials in the relative position of individuals of varying x, according to (2) and (3):

$$E(z_1|x+1) - E(z_1|x) = \frac{\rho}{\sigma_{y_1}} = \frac{\rho}{\sqrt{\rho^2\sigma_x^2 + \sigma_{\varepsilon_1}^2}}$$

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

Consider the simplest case where there are no direct effects of the explanatory variables ($\beta = 0$), no carryover effects of previous inequalities ($\theta = 0$), and scores at *t*=1 and *t*=2 follow the same metrics ($\omega = 1$):

$$[E(z_2|x+1) - E(z_2|x)] - [E(z_1|x+1) - E(z_1|x)] = \frac{\rho}{\sqrt{\rho^2\sigma_x^2+\sigma_{\varepsilon_1}^2+\sigma_{\varepsilon_2}^2}} - \frac{\rho}{\sqrt{\rho^2\sigma_x^2+\sigma_{\varepsilon_1}^2}} < 0$$

Here the average distance between children of different backgrounds narrows, simply because at *t*=2 there is higher (unexplained) variability. Yet, the same reduction could arise for completely different reasons; for instance, if given previous scores children of disadvantage backgrounds perform better ($\beta < 0$) or if achievement growth is negatively related to performance at *t*=1 ($\theta < 0$).

Summing up, diff-in-diff strategies relying on cross-sectional data do not allow to distinguish between the relevant sources of the observed changes in the position of groups of pupils relative to each other. To do so we need to estimate the dynamic model (3) directly.[10]

4. Estimation of the dynamic model

The problem we address here is how to estimate (3) in the absence of genuine longitudinal data, where the data derive from independent cross-sectional surveys held at different stages of schooling, each being a random sample of the same underlying population of children.

Consider individuals interviewed at *t*=2 (CS2): even if their own lagged scores y_1 are unobserved, we can obtain y_1 for different but "similar" children – i.e. sharing the same observed

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

[10] In this perspective, we may question the validity of the diff-in-diff approach employed in Waldinger (2007) and Jakubowski (2010) to evaluate the effect of tracking on mean achievement and socioeconomic inequalities.

characteristics – by exploiting individuals interviewed at *t*=1 (CS1). A simple strategy would be to randomly select for each child in CS2 a similar child in CS1, and use her score y'_1 instead of true y_1. This strategy, however, leads to severely biased results, because the lagged score is affected by (large) measurement error. Conventional methods to correct for measurement error (Fuller, Hidiroglou; 1978) are not appropriate here, because they assume the CEV (*classical error in variables*) condition, i.e. that measurement error is independent of true values. In this case, however, the error is related to *both* true and observed values. In fact, if $y_{i1} = \mu_i + u_{i1}$ and $y'_{i1} = \mu_i + u'_{i1}$ (where μ_i is the mean score of an individual with given x), measurement error is $y_{i1} - y'_{i1} = u_{i1} - u'_{i1}$.

An alternative strategy could be the estimation of a regression model for cell means, where cells are defined as groups of similar individuals. In this case, instead of matching individuals from different cross-sections, we match cells, i.e. groups of children sharing the same characteristics.[11] The main advantage of this strategy over individual matching is that measurement error of group means is smaller. If the number of cells is fixed, the sampling variance of the cell means tend to zero as the overall sample size expands. Hence, OLS estimates are consistent. In any particular sample, however, the presence of measurement error in y_1 will lead to biased estimates. Again, the method proposed by Fuller, Hidiroglou (1978) does not solve the problem, because the error is not CEV and therefore the correlation between the error term and lagged performance score is not fully eliminated.

Note also that if the explanatory variables are the same at *t*=1 and *t*=2, the model for *t*=2 is unidentified because y_1 cell means are a linear function of x. Therefore, as we will also argue for the imputed regression technique described below, we need to find a variable affecting y_1 but not y_2, and define cells by taking this auxiliary variable into account.

4.1 Imputed regression

The conditions for identification and consistent estimation of general linear dynamic panel data models with repeated cross-sections are discussed in Moffitt (1993) and later developed by Verbeek, Vella (2005). The basic idea is that the lagged dependent variable can be replaced by a predicted value from an auxiliary regression using individuals observed in previous cross-sections: the resulting measurement error will generally be uncorrelated with estimated lagged performance and therefore will not lead to inconsistent estimates, as is the case with CEV errors. Measurement error, however, must be uncorrelated also with all other explanatory variables. Whether these

[11] Cell mean regression has been applied in a variety of contexts to analyze repeated cross-sectional data. Card and Lemieux (1996), for example, use it to analyze changes in returns to skill and wage inequalities.

conditions are met depends crucially on the nature of the dynamic model and of the model employed to predict lagged values. Verbeek and Vella (2005) argue that these requirements are unrealistic in many contexts; they show, however, that they hold if there are no time-varying exogenous variables or the time-varying exogenous variables are not auto-correlated. Our context is particularly simple: in first place, because the only source of dynamics in the process is the autoregressive component, while individual fixed effects are not included; secondly, because the explanatory variables of interest are all time-invariant socio-demographic characteristics.

Yet, if the set of independent variables is identical for y_1 and y_2 – a likely occurrence when we focus on performance differentials across ascribed individual characteristics – model (3) is unidentified when substituting y_1 with $\hat{y}_1$. Hence, in order to bypass collinearity, we must find a variable w affecting performance at t=1 but not directly related to later performance.

Assuming the following model for y_1:

$$y_{i1} = \mu_1 + \rho x_i + \delta w_i + \varepsilon_{1i} \tag{6}$$

we substitute y_1 with its OLS estimate $\hat{y}_1'$ derived from CS1. Expressed in terms of $\hat{y}_1'$ the model then becomes:

$$y_{i2} = \mu_2 + \gamma \hat{y}_{i1}' + \beta x_i + [\gamma(y_{i1} - \hat{y}_{i1}') + \varepsilon_{2i}] \tag{7}$$

Since y and y_1' are independent draws from the same population, in large samples $\hat{y}_1$ and $\hat{y}_1'$ are nearly coincident. Thus the estimation of (7) is nearly equivalent to the estimation of $y_{i2} = \mu + \gamma \hat{y}_{i1} + \beta x_i + [\gamma(y_{i1} - \hat{y}_{i1}) + \varepsilon_{i2}]$. As seen above, measurement error derived from using $\hat{y}_1$ instead of true y_1 is not CEV: however, for OLS properties $(y_1 - \hat{y}_1)$ is uncorrelated with $\hat{y}_1$. Thus, OLS estimates of (7) are consistent.[12] However, the resulting standard errors are larger than with longitudinal data.

In Appendix A we describe a simulation study designed to evaluate the bias associated with the matching strategies outlined above and to provide an order of magnitude of the standard errors obtained with imputed regression. As for the latter, the main result is that standard errors are largely inflated, and their magnitude strongly depends on sample size and on the predictive power of w.[13]

The imputed regression strategy has an advantage over genuine panel data models: by explicitly addressing the issue of measurement error, it provides consistent estimates also if test scores are imperfect measures of achievement. Let observed scores $y_t = y_t^* + v_t$, with true scores y_t^*

[12] In principle, using $\hat{y}_1'$ instead of $\hat{y}_1$induces a small correlation between the error term and explanatory variables. Let us further inspect (7): $y_{i2} = \mu + \gamma \hat{y}_{i1}' + \beta x_i + \gamma\big((y_{i1} - \hat{y}_{i1}) + (\hat{y}_{i1} - \hat{y}_{i1}')\big) + \varepsilon_{2i}$. The term $(\hat{y}_{i1} - \hat{y}_{i1}')$ is not totally independent of x and $\hat{y}_{i1}'$; however, with reasonable sample size it accounts for a negligible share of the total error term. This caveat, therefore, has no relevant practical implications.

[13] If x explains a large portion of the variance of y_1 while w does not, residuals might be small, but $\hat{y}_1$and x will be nearly collinear and the resulting standard errors of the estimates large.

independent of measurement error v_t. The estimation of equation (2) poses no problems, as measurement error affects the dependent variable. As for model (3), consider the equation for true scores: $y_{i2}^* = \mu_2 + \gamma y_{i1}^* + \beta x_i + \varepsilon_{i2}^*$. The equation for observed scores when the predicted value of y_1 is introduced is $y_{i2} = \mu_2 + \gamma \hat{y}_{i1} + \beta x_i + \varepsilon_{i2}^* + v_{i2} + \gamma(y_{i1}^* - \hat{y}_{i1})$. The composite error term is independent of all explanatory variables: in particular, $(y_1^* - \hat{y}_1)$ is independent of $\hat{y}_1$ because the regression coefficients of (2) are unbiased in spite of measurement error. Thus, aside from the effect of sampling variability, predicted lagged values are the same if estimated on true or observed scores.

Choice of the variables allowing identification

Our main aim is to estimate consistently model (3). We may therefore use equation (2) to predict y_1 as precisely as possible, regardless of the nature of explanatory variables. In this perspective, however, two conditions are necessary:

(i) Additional predictors w cannot be relevant variables for achievement at *t*=2.

In other words, they must be valid instruments. For example, assume there are two or more indicators of family background, each capturing different features that affect learning throughout schooling life. If in the attempt to avoid collinearity we exclude either one from the model for y_2, we get biased estimates, because the omitted variable, entering the error term in equation (7), is correlated with $\hat{y}_1$.

(ii) Additional predictors w must be observed at both cross-sections CS1 and CS2

since $\hat{y}_1$ is introduced in the model for y_2 for given x and w. As a consequence, natural candidates such as school characteristics at *t*=1 – which could be good instruments as they are liable to affect current but not future performance – cannot be employed, because school features at *t*=1 are usually not recorded in CS2.

It is therefore difficult to find an appropriate instrument. In our empirical analysis we will use the month of birth, as there is evidence that younger children perform more poorly than their older peers, and that later achievement does not depend on age given previous achievement.

4.3 Limitations of pseudo-panel estimation

The focus of this paper is the development of learning inequalities over different stages of the educational career: in this perspective the explanatory variables of interest are time-invariant socio-demographic factors. However, with genuine longitudinal data a variety of challenging questions involving the effect of potentially endogenous variables (such as school characteristics) can be

addressed employing value-added models. Would pseudo-panel modeling allow to tackle these issues? Unfortunately, the general answer is negative.

Consider the following model, where s_2 are school characteristics at *t*=2:

$$y_{i2} = \mu_2 + \gamma y_{i1} + \beta x_i + \pi s_{i2} + \varepsilon_{i2}.$$

If we substitute y_1 with its OLS estimate $\hat{y}_{i1} = \hat{\mu}_1 + \hat{\rho} x_i + \hat{\delta} w_i$ from CS1, we obtain:

$$y_{i2} = \mu_2 + \gamma \hat{y}_{i1} + \beta x_i + \pi s_{i2} + (\gamma(y_{i1} - \hat{y}_{i1}) + \varepsilon_{i2}) \quad (8)$$

The problem is that school features at *t*=2 are typically not independent of $(y_1 - \hat{y}_1)$, as higher ability children are more likely to choose "better" schools. This establishes a correlation between s_2 and the error term. To obtain consistent estimates of (8) we need to consistently estimate $E(y_1|x, w, s_2)$, and to do so, school characteristics at *t*=2 should be observed also at *t*=1, which is clearly impossible.[14,15]

5. Inequalities in Italy at the turning point between primary and lower secondary school

5.1 Italian schooling system and data

In the Italian educational system children enter school at age 6 and follow an eight year period of comprehensive schooling: primary education, lasting five years, and lower secondary education, lasting three. Lower secondary school ends with a nationally-based examination, after which students choose between a variety of upper secondary educational programs, broadly classified into academic, technical and vocational tracks. Despite the absence of performance-related admission restrictions, the academic track is much more demanding than the other tracks.

The study of inequalities at the transition between primary and lower secondary school is of particular interest because the emphasis on achievement requirements greatly increases, as lower secondary school is perceived as the period in which children get prepared for upper secondary education. A similar research question is addressed in De Simone (2013), who focuses on the time-span between grade 4 and 8 with TIMSS.[16]

[14] In most European countries between age 10 and 16 children are tracked into academic and vocational programs, usually offered by distinct institutions. The endogeneity problem described above would be particularly severe if the first assessment was held before tracking and the second one after, because track choice is strongly related to ability.

[15] The same argument applies to the investigation of the effects of children's behavior at *t*=2 (e.g. effort, time for homework), as behavior at *t*=2 is likely to be dependent on achievement at *t*=1.

[16] Incidentally, aside from our own work (previous version in Contini and Grand, 2012), De Simone (2013) is the only other contribution in the literature that we are aware of using pseudo-panel modeling to study achievement inequalities. As shown by the simulations described in the Appendix, however, TIMSS does not have a sufficient sample size to ensure reliable results. Our empirical work also differs from De Simone (2013) in the choice of explanatory variables and identification strategy.

We use the repeated cross-sectional data of the standardized learning assessment administered by the Italian National Evaluation Institute (INVALSI[17]) to the entire student population of 5th (end of primary school) and 6th graders (lower secondary school), consisting of approximately 500,000 students per grade (INVALSI, 2011). We "link" the survey administered in 2010 to 5th graders to the survey administered in 2011 to 6th graders, following children born in 1999.

Tests cover the domains of reading and mathematics, and were designed following the experience of international assessments. Students are asked to fill a questionnaire recording personal information, including family composition and home possessions, while school boards provide information on parental background. School teachers are normally in charge of test administration; however, in order to control for cheating, a random sample of classes (of about 30,000-40,000 students) take the tests under the supervision of personnel external to the school. This sample represents a benchmark to evaluate performance scores of the general population. Sample mean scores are generally smaller than population values; the differential is interpreted as evidence of cheating (Quintano et al., 2009). Due to its better quality, our empirical analyses will be based on this sample data.

Table 1. Descriptives. 5th grade (2010 assessment) and 6th grade (2011 assessment)

DEPENDENT VARIABLES					
VARIABLE	DESCRIPTION	MEAN 5TH	S.D. 5TH	MEAN 6TH	S.D. 6TH
Score reading	Percentage correct answers reading test	67	18	63	17
Score math	Percentage correct answers math test	62	18	47	19
EXPLANATORY VARIABLES					
VARIABLE	DESCRIPTION	MEAN 5TH	S.D. 5TH	MEAN 6TH	S.D. 6TH
Female	Percentage females	0.49		0.49	
Books	N° of books at home*	1.97	1.20	2.08	1.23
ESCS	Economic-Socio-Cultural Status Index	-0.02	1.00	0.03	1.00
Native	Native	0.91		0.90	
2G	Second-generation migrant	0.04		0.04	
1G	First-generation migrant	0.05		0.06	
North-West	Living in North-West	0.24		0.24	
North-East	Living in North-East	0.17		0.17	
Centre	Living in Centre	0.18		0.18	
South	Living in South	0.24		0.24	
Islands	Living in the Islands	0.17		0.17	
SAMPLE SIZE	READING-MATH	33997-33530		37196-37183	

* 0=0-10 books; 1=11-25 books; 2=26-100 books; 3=101-200 books;4=>200 books

NOTES. The percentage of natives, first and second-generation immigrants are computed over the entire sample.
All other descriptives refer to natives and second-generation immigrants only.
Standard deviations for binary variables are not reported.

[17] Istituto Nazionale per la Valutazione del Sistema educativo di Istruzione e formazione.

Performance is measured by the percentage of correct answers, varying between 0 and 100. Scores are not vertically equated, so achievement is not comparable across grades. We employ two measures of socioeconomic status. The first is the number of books at home, in line with contributions in the economics of education.[18] The second is the standardized index ESCS (*Index of Economic-Socio-Cultural Status*) provided by the National Agency and derived from data on home possessions, parental education and occupation.[19] We also investigate gender, immigrant background and territorial differentials (according to macro-areas: North-West, North-East, Centre, South, Islands). All variables are summarized in Table 1.

5.2 Empirical analysis on native children

In this section we run pseudo-panel models on native children. The inclusion of immigrant students is problematic because they experience grade repetition much more frequently than natives. As will be discussed in the next section, grade repetition at *t*=1 may represent a threat to consistent estimation. Yet, since less than 1% repeat the school year in primary school, this issue is almost irrelevant for native children.[20]

Our instrumental variable is the month of birth, as there is empirical evidence that younger children perform more poorly than their older peers. The first identifying assumption, verified within CS1, is that the month of birth affects y_1. The second one is that the month of birth does not directly affect y_2: $E(Y_2|y_1, x, w) = E(Y_2|y_1, x)$; since this assumption cannot be evaluated without longitudinal data, it was tested on a different dataset.[21]

Results for both reading and math are summarized in Table 2. The first columns contain the estimates of cross-sectional model (6) for y_1; the second report the estimates of the cross-sectional model $y_{i2} = \mu_2 + \beta x_i + \varepsilon_{i2}$; the third contain the estimates of dynamic model (7).

[18] Children are asked to select a picture depicting a variable number of shelves with books.
[19] This measure, drawn from the international survey PISA, is the first factor of a principal component analysis.
[20] Models are estimated on children of birth cohort 1999 (the regular birth cohort for grade 5 in 2010 and grade 6 in 2011). In addition to children repeating grades, we exclude children born in 2000 (approximately 9% of the sample). This exclusion might cause non-random sample selection, as children starting school at age 5 (before the regular age) instead of age 6, could be more mature than their peers or have higher innate ability. For this reason we carried out some robustness checks by including these children in the estimation, attributing to them either month of birth 12 (as if they were all born in December 1999), or month 13 (for those born in January 2000), month 14 (for those born in February 2000) and so on. Since we find only minor changes, these results are not shown.
[21] This dataset was collected in 2010 within the project "Scacchi e Apprendimento della Matematica" - *Chess and Math Learning* (Argentin and Romano "Standing on the Shoulders of Chess Masters: Using RTCs to Evaluate the Effects of Including Chess in the Italian Primary School Curriculum", in Besharov D. (eds), *Evaluating Education Reforms: Lessons from Around the Globe*, Oxford University Press, forthcoming). It provides two repeated performance measures at the beginning and the end of third grade. We thank the authors for permission to use these data.

Table 2. Estimates of cross-sectional and pseudo-panel data models. Native children.

	READING			MATHEMATICS		
Variables	5th grade cross-section	6th grade cross-section	6th grade dynamic	5th grade cross-section	6th grade cross-section	6th grade dynamic
Costant	65.8***	58.7***	14.1*	61.5***	45.3***	-21.1***
Month[1]	-0.3***			-0.3***		
Books[2]	2.4***	2.1***	0.5	2.2***	2.5***	0.1
ESCS	3.3***	3.9***	1.6***	3.0***	3.7***	0.4
Female	0.8**	2.2***	1.7***	-3.1***	-3.8***	-0.3
North East	-0.5	-0.6	-0.3	-1.3*	1.0	2.4***
Centre	-2.6***	-1.4***	0.5	-2.2**	-2.9***	-0.4
South	-3.6***	-2.5***	0.1	-0.9	-5.2***	-4.0***
Islands	-6.3***	-6.5***	-1.9*	-4.0***	-8.4***	-3.7***
y_1			0.702***			1.116***
R^2	0.132	0.158	0.160	0.091	0.156	0.160
sample size	26616	29637	29637	27333	29636	29636

*p_value<0.05; **p-value<0.01; ***p-value<0.005
NOTES. The estimation has been performed only on children born in 1999.
1. January=1,… December=12
2. 0=0-10 books; 1=11-25 books; 2=26-100 books; 3=101-200 books;4=>200 books
Cluster robust standard errors (Huber-White estimator, run with STATA). Clusters defined by classes.

Looking at cross-sectional results, the effects of individual ascribed characteristics are substantial at both assessments. In line with the international literature, socioeconomic status emerges as a strong predictor of performance: the coefficients of both indexes – the number of books at home and ESCS – are large and highly statistically significant. Consider the reading assessment in 5th grade: a unit increase in the number of books variable yields an average increase of 2.4 points; a unit increase in ESCS yields an increase of 3.3 points. This implies that when comparing a child with the lowest category of number of books (0) and a low value of ESCS (-2) with a child with the highest category of number of books (4) and a high value of ESCS (+2), on average the latter scores 22.8 points more than the former. Females perform better than males in reading and worse in mathematics. Territorial differentials are dramatic, in particular along the North-South divide. For example, other things being equal, at both grades 5 and 6 the average reading score in the Islands is more than 6 points lower than in the North-West. The achievement level in mathematics does not differ significantly between the South and the North in grade 5 (while it is much poorer in the Islands), but in grade 6 the divide appears to widen substantially.

We now turn to the interpretation of the results of pseudo-panel estimates. β coefficients in the dynamic model measure the extent to which achievement growth between t=1 and t=2 differs across x-categories, when comparing two equally performing children at t=1. We observe substantive socioeconomic and gender effects in reading but not in math. As for reading, children from high socioeconomic background, already advantaged in grade 5, do better in grade 6 than previously equally performing children of lower backgrounds (this is evident from the coefficient of ESCS,

while that of books at home is not significant). On the other hand, they do not do any better or worse in mathematics. Similar results hold for gender effects: girls improve relative to boys in reading, while their disadvantage in math does not develop further. The opposite finding holds for area effects: they are small in reading but very large in math. The most noticeable result is that, given 5th grade achievement, 6th graders living in the North largely outperform their Southern peers. On the other hand, children living in the North-East close the math achievement gap with the North-West and perform best given previous scores.

Goodness of fit (measured by R-squared) is quite low in all models. This result is hardly surprising for cross-sectional models: indeed, ascribed characteristics cannot explain a large portion of the variability, even in educational systems with large inequalities across children of different backgrounds. However, also the R-squared of the dynamic model is low; adding predicted lagged scores does not yield a substantial increase in goodness of fit, because measurement error due to the substitution of true lagged performance with an estimated value is (very) large.

Let us now relate our findings with the theoretical arguments exposed in section 4. We have argued that cross-sectional regression coefficients are not fully informative on the development of inequalities between the two assessments. Consider cross-sectional estimates at grades 5 and 6 and take their difference for each explanatory variable. Their difference (corresponding to diff-in-diff) diverges from dynamic coefficients. As shown by equation (5), diff-in-diff and dynamic coefficients coincide if $\omega = 1$ and $\theta = 0$. In this case the coefficient of the lagged score in (3) is $\gamma = 1$. Since $\hat{\gamma}$ is farther from 1 in reading than in math, the divergence is larger for the former. If we were to interpret the differential between the North-West and the Islands from cross-sectional coefficients (-6.3 in grade 5, -6.5 in grade 6), we would conclude that this inequality has hardly changed between 5th and 6th grade. Instead, the lagged performance coefficient in the dynamic model (-1.9) clearly indicates that the disadvantage of the latter has substantially widened. Similarly, from cross-sectional point estimates (3.3 in grade 5, 3.9 in grade 6) we would conclude in favor of a small ESCS effect, whereas the estimate of the dynamic model (1.6) points to considerable increasing socioeconomic background inequality.

On the other hand, dynamic regression coefficients measure achievement growth differentials across x-categories when comparing two equally performing children at *t*=1. Thus, pseudo-panel βs capture inequalities developed in the time span between the two surveys that can be directly ascribed to each explanatory variable. As shown in (5), however, if individual growth is affected by previous performance, growth differentials between x-categories are also driven by preexisting performance gaps. Yet, this component cannot be evaluated, because the effect of previous performance on growth is unidentified without vertically equated scores.

5.3 Children repeating grades

In a typical panel data setting y_1 is the value of y at calendar time 1, y_2 is the value of y at time 2 and so on. Instead, in educational surveys where children of specific *grades* are interviewed, y_1 represents performance score at a specific grade (here, 5th grade) and y_2 the score at a later grade (here, 6th grade). As a consequence, if some children are required to repeat a grade, there will be two y values at this grade. This poses no particular problems with genuine panel data[22], but the situation is more complex for pseudo-panel estimation.

Figure 2 may help to illustrate the problem. In the columns we indicate calendar time. Consider the 5th grade assessment held at calendar time *T* and the 6th grade assessment held at *T*+1: pupils participating in these surveys belong either to cohort *k* (regulars) or to cohort *k*-1 (one year late). In the rows we describe possible paths. Rows 1-3 refer to children of birth cohort *k*: with a regular career, failing grade 6 and failing grade 5. Rows 4-5 depict children of cohort *k*-1 repeating either grade 5 or grade 6. To simplify the picture, we assume that children may fail only once[23] and that no repetitions occur before grade 5.[24]

Figure 2. Birth cohorts and assessments.

T-1	*T*	*T*+1	*T*+2	Row	Fail	Cohort
	5 regular (birth cohort *k*)	6 regular (birth cohort *k*)	7 regular (birth cohort *k*)	1	-	*k*
	5 regular (birth cohort *k*)	6 regular (birth cohort *k*)	6 one year late (birth cohort *k*)	2	Grade 6	*k*
	5 regular (birth cohort *k*)	5 one year late (birth cohort *k*)	6 one year late (birth cohort *k*)	3	Grade 5	*k*
5 regular (birth cohort *k*-1)	5 one year late (birth cohort *k*-1)	6 one year late (birth cohort *k*-1)		4	Grade 5	*k*-1
5 regular (birth cohort *k*-1)	6 regular (birth cohort *k*-1)	6 one year late (birth cohort *k*-1)		5	Grade 6	*k*-1

Let us focus on children of birth cohort *k* and assume that we are interested in first time pupils attend each grade (underlined). We now argue that if there are no children failing before grade 5 the estimates based on regular children (grey shadowed cells) are unbiased. Consider equation (6), the cross-sectional model for y_1. The relevant set for its estimation consists of pupils of cohort *k* participating to the 5th grade assessment for the first time at time *T*. These children are all observed.

[22] The issue would be whether to model y_2 as dependent on the first or the second y_1.
[23] Repeated failures account for a negligible share of children in our data.
[24] The INVALSI survey does not provide information on repetitions but only the year of birth; therefore we cannot distinguish between children repeating the current grade or previous grades. We assume that children of earlier birth cohorts in grade 5 are repeating grade 5; the reason is that grade 5 represents the transition point from primary school, in which emphasis on achievement is low, to secondary school, in which it is much higher.

Consider now the estimation of (7), the dynamic model for y_2. Children in rows 1 and 2 pose no problem, since they participate to the 6th grade assessment for the first time at T+1. Children in row 3 (repeating grade 5), instead, participate at T+2; hence, they are excluded from the estimation. This exclusion is related to achievement at t=1, which is an explanatory variable in model (7); therefore, it does not have severe consequences on the estimates, because sample selection on independent variables inflates standard errors but does not lead to biased estimates.[25]

Could we include children of birth cohort k-1 repeating grades to replace the unobserved children in row 3? Children in row 4 (repeating grade 5) are homologous to children in row 3; however, children in row 5 (repeating grade 6) are homologous to children in row 2, which are already included in the estimation. Since previous school history is not recorded, it is not possible to distinguish children of row 4 and row 5. Hence, we may either include or exclude both of them. Including both of them is like duplicating some observations and leads to biased estimates, as shown in the simulation exercise described in Appendix A (Table A4).

5.4 Including children with an immigrant background

The inclusion of immigrant background students in the analyses and the evaluation of ethnic educational inequalities is problematic. First-generation immigrants are often placed in earlier grades at arrival in Italy, and therefore our instrument, children's age, would be endogenous. Moreover, this population is subject to significant changes in the short run due to territorial mobility. Newly arrived immigrants are also likely to have severe language problems and might therefore be excluded from the assessment. For these reasons, we exclude first-generation immigrants from empirical analyses.[26]

Second-generation immigrants are less subject to population instability, have been entirely exposed to the Italian schooling system and usually enter school at the regular age. Therefore comparing their achievement to that of natives is meaningful. In Italy, however, poorly achieving children are often required to repeat a grade. The repetition probability increases over the schooling career for all students and it is generally much higher for children with an immigrant background. According to the INVALSI data, in 5th grade (survey 2010) less than 1% of natives and around 6% of second-generation immigrants were older than the regular age, while in 6th grade (survey 2011) the proportion was 5% for natives and 15% for immigrants.

[25] Limiting the analysis to regular children would produce biased estimates if year failure also occurred before the grade attended at t=1. In this case children failing before this grade would be excluded from the estimation of the model for y_1, because at calendar time T they would still be attending an earlier grade. Regular children would represent a positively selected sample on the *dependent* variable, leading to biased estimates of $E(y_1)$ and of the coefficients of (7).

[26] We define first-generation immigrants as children born abroad to two foreign-born parent and second-generation immigrants children born in Italy to two foreign-born parents.

As already pointed out, these figures do not allow to distinguish between current and previous repetitions. Consistently with the discussion of previous section, we assume that older children failed grade 5: under this condition the estimates based on children born in 1999 are unbiased. Results are shown in Appendix B (Table B.1).

Cross-sectional estimates indicate that in the reading assessment children of immigrant background perform worse than natives by 4-6 points and in grade 6 the advantage of immigrant girls over boys is significantly larger than among natives. As for mathematics, in grade 6 there are significant territorial differences. In the North and the Centre, immigrant boys perform worse than native boys by approximately 4 points in both assessments; instead, immigrant girls do only marginally worse than native girls in grade 6. In the South and the Islands the immigrant-native performance gap is substantially smaller.

The estimates of the dynamic model reveal that, given achievement in grade 5, in grade 6 immigrant girls do not differ significantly from native girls in reading, while they substantially improve in math. Instead, immigrant boys do worse than native males in reading and remain stable in math. On math scores we also observe interaction effects between immigrant status and area of residence: while among natives the North-South divide widens between grade 5 and 6 (see also Table 2), among second-generation immigrants we observe no relevant changes. Overall, these findings suggest that although second-generation immigrant students are on average lower performing than natives, their disadvantage is largely established by the end of primary school. There is no evidence of growing immigrant background inequality in lower secondary school for girls, whereas for boys inequality widens in reading skills but not in mathematics.

6. Summary and conclusions

In this article we discuss the estimation of dynamic models from repeated cross-sectional standardized learning surveys, with the aim to assess how inequalities related to ascribed individual characteristics develop over childhood. Drawing from Verbeek and Vella (2005), we propose an imputed regression strategy allowing to "link" two surveys; the basic idea is that lagged scores are replaced by predicted values derived from a regression on the previous cross-section. We show that – given our research question and with appropriate explanatory variables – this strategy delivers consistent estimates of the parameters of interest. Moreover, by explicitly addressing the issue of measurement error, imputed regression provides consistent estimates of the parameters of interest even with test scores imperfectly measuring achievement.

If repeatedly applied to different segments of the schooling career, the method allows to investigate how inequalities develop over the educational life course in contexts where longitudinal

data are not available. Moreover, by exploiting international surveys, we could analyze cross-country differences and explore the relationship between the development of inequalities and features of the educational systems.[27] This would represent a significant contribution to the literature on the effect of institutions on achievement inequalities, moving the focus from a static to a dynamic perspective.

The main drawback, however, is that due to the substitution of true lagged scores with an estimate, standard errors are largely inflated. This result is shown in the simulation exercise described in Appendix A. The practical implication is that sizable samples are needed. This conclusion limits the applicability of the proposed strategy: to date, it would be difficult to exploit international assessments for cross-country comparisons, as samples are usually not large enough.[28]

We apply the method to the large-scale learning assessments on reading and math literacy, carried out by the Italian National Evaluation Agency in 2010 and 2011 on children attending 5th and 6th grades. We evaluate how gender, socioeconomic, immigrant background and territorial inequalities develop at the transition between primary and lower secondary school. The empirical analysis reveals that gender and socioeconomic inequalities widen in reading literacy but remain stable in math. On the contrary, the North-South divide does not change in reading but severely increases in math: this result suggests that math teaching in lower secondary school could be much less effective in the South and Islands. Immigrant background differentials are largely established at the end of primary school; with the exception of reading skills for boys, second-generation immigrants do not lose ground with respect to natives in grade 6, and girls even catch up part of their disadvantage in math.

[27] TIMSS could be particularly well suited for this purpose, as it administers tests to children of given birth cohorts at grades 4 and 8.

[28] Additional instruments could help increasing the efficiency of the estimation, but as we have argued above, they are difficult to find with the information currently available. School characteristics at t=1 might represent valid instruments, but are not recorded at t=2.

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

Betebenner, D.W. and Linn, R. (2010) Growth in student achievement: issues of measurement, longitudinal data analysis, and accountability, Working Paper of the Centre for K-12 Assessment & Performance Management

Card, D. and Lemieux, T. (1996) Wage dispersion, returns to skill, and black-white wage differentials, *Journal of Econometrics*, **74**, 319-361

Contini, D. and Grand, E. (2012) Estimating student learning value-added models from repeated cross-sections. *Proceedings of the Congress of the Italian Statistical Society*, Rome 6-8 June 2012, (Available from http://meetings.sis-statistica.org/index.php/sm/sm2012/paper/viewFile/2099/134)

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* (eds E. Hanushek and F. Welch), Chapter 12, pp. 697-812. Amsterdam: North Holland.

Deaton, A. (1985) Panel data from time series of cross-sections, *Journal of Econometrics*, **30**, 109-126.

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

Fuller, W.A. and Hidiroglou, M.A. (1978) Regression estimation after correcting for attenuation, *Journal of the American Statistical Association*, **73**, 99-104

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.

INVALSI (2012) *Le prove INVALSI 2011. Rapporto tecnico sulle caratteristiche delle prove INVALSI* 2011. Rome: INVALSI. (Available from http://www.invalsi.it/areadati/SNV/10-11/Rapporto_tecnico_prove_invalsi_2011.pdf).

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

Jerrim, J. and Vignoles, A. (2012) Social Mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes, *Journal of the Royal Statistical Society*: Series A (Statistics in Society). doi: 10.1111/j.1467-985X.2012.01072.x.

Moffitt, R. (1993) Identification and estimation of dynamic models with a time series of repeated cross-sections, *Journal of Econometrics*, **59**, 99-123

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

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. (Available from http://dx.doi.org/10.1787/9789264091450-en)

OECD (2010b) PISA 2009 results: overcoming social background. Equity in learning opportunities and outcomes. Volume II. (Available from http://dx.doi.org/10.1787/9789264091504-en).

Quintano, C., Castellano, R. and Longobardi, S. (2009) A Fuzzy Clustering Approach to Improve the Accuracy of Italian Student Data. An Experimental Procedure to Correct the Impact of Outliers on Assessment Test Scores. *Statistica & Applicazioni* **7**,149-171.

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.

Verbeek, M. and Vella, F. (2005) Estimating dynamic models from repeated cross-sections, *Journal of Econometrics*, **127**, 83-102

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

Appendix

Appendix A. Simulation study

Imputed regression: varying sample size and predictive power of the instrument

In section 4.1 we demonstrated that the imputed regression strategy yields unbiased estimates of regression coefficients, but standard errors are going to be inflated with respect to the estimation on genuine panel data, because the error term incorporates measurement error due to the substitution of true lagged scores with an estimate. In an attempt to assess the practical relevance of the imputed regression strategy in educational achievement surveys, we run a simulation study to explore the behavior of the estimates with changing sample size and predictive power of the instrument w.

For each replication, we first generate values of y_1 according to model (2), and then generate y_2 according to model (3). Parameters are set approximately at the estimated values in our empirical analysis.[29] We consider two alternative sample sizes: 5000 (a typical size in the international assessments TIMSS) and 30000 (the sample size in the Italian survey). The column representing our case-study is grey shadowed. We run 1000 replications, and then compute the average value of the estimates and two statistics related to standard errors: the standard deviation of regression coefficient estimates across replications (se.1), and the mean value of the resulting standard error estimates within each replication (se.2). Results are summarized in Table A1. As we can see, the regression coefficient estimates are on average very similar to true parameters. Standard errors are generally large, in particular for small n and coefficient of w. As these parameters increase, the estimates become more precise.[30]

We can draw a major conclusion: in order to obtain reliable estimates large samples and good instruments are needed. With 5000 individuals, a typical country-level sample size in international assessment, the estimates are quite unstable (they can be totally unreliable with a very poor instrument). Instead, with a sample of 30000 students results can be considered reliable enough.

[29] The standard deviation of the error term in (3) cannot be estimated with pseudo-panel, so we have to guess. Since model (3) is conditional on previous performance, the error term variability is likely to be smaller than in (2). In the empirical analysis the estimate of the coefficient of w is -0.3; in the simulation we let it vary between -0.1 and -0.5.

[30] Se.1 always exceeds se.2 (although the differences are relatively small, in particular for large samples and predictive power of w). The reason is that the estimates of regression coefficients of y_2 are based on predicted values of y_1; the latter vary across replications, as also CS1 is a random sample. Se.1 incorporates this source of variability while within regression estimates of standard errors neglect it, because they are conditional on $\hat{y}_1$. To confirm this argument we carried out additional simulations where we plugged true $E(y_1 | x, w)$ into equation (7) instead of $\hat{y}_1$: resulting standard deviations of the estimates were nearly identical to average estimates of standard errors.

Table A1. Imputed regression. Varying sample size and coefficient of the month of birth.

	TRUE VALUE	N=30000 β_{month}=-0.1	N=5000 β_{month}=-0.1	N=30000 β_{month}=-0.3	N=5000 β_{month}=-0.3	N=30000 β_{month}=-0.5	N=5000 β_{month}=-0.5
SES	2	1.79 (1.80,1.21)	1.96 (32.1,4.9)	1.99 (0.45,0.37)	1.85 (1.23,0.95)	1.99 (0.27,0.23)	1.96 (0.68,0.58)
Sex	2	1.85 (1.35,0.92)	2.16 (35.1,5.2)	1.99 (0.39,0.34)	1.88 (1.11,0.85)	1.98 (0.31,0.25)	1.97 (0.74,0.63)
Area	2	-1.90 (0.90,0.61)	-1.81 (29.7,4.8)	-2.00 (0.23,0.20)	-1.93 (0.63,0.50)	-1.99 (0.16,0.13)	-1.99 (0.37,0.32)
*y*1	0.7	0.75 (0.45,0.30)	0.78 (15.2,2.4)	0.70 (0.11,0.09)	0.74 (0.30,0.23)	0.70 (0.07,0.06)	0.71 (0.16,0.14)
Const	20	17.44 (22.20,14.90)	15.68 (741.8,119.8)	19.86 (5.26,4.43)	18.21 (14.52,11.28)	19.76 (3.19,2.59)	19.57 (7.65,6.48)

NOTES.
Models generating data:
$y_1 = 50 - \beta_{month} w + 4SES + 3sex - 2area + \varepsilon_1;\ y_2 = 20 + 0.7y_1 + 2SES + 2sex - 2area + \varepsilon_2$
$\sigma(\varepsilon_1) = 16, \sigma(\varepsilon_2) = 12$
Range of explanatory variables*: month of birth* (1-12); *SES*(1-5); *sex*(0-1); *area*(1-4)
Average over 1000 replications' estimates.
In parenthesis: (st. dev of estimates over replications, mean se. of the estimates).

Individual matching, cell mean regression and imputed regression

In this exercise we compare the estimates obtained with different estimation methods: individual matching, cell regression and imputed regression. Sample size is 30000. Individual matching is performed by substituting true y_1 with a random value drawn from the same distribution given explanatory variables. As for cell mean regression, we define cells according to the discrete values of all explanatory variables, obtaining 600 cells, with an approximate size of 50 units each.

Table A2. Comparison of alternative estimation strategies. Individual matching, cell mean regression, and imputed regression.

	TRUE VALUE	Individual matching (1)	Cell matching (1) 600 cells	Cell matching (2) 600 cells	Cell matching (3) 600 cells	Imputed regression (1)
SES	2	4.79 (0.07.0.07)	4.33 (0.17.0.17)	2.83 (0.13, 0.13)	2.26 (0.10, 0.10)	1.99 (0.45.0.37)
Sex	2	4.08 (0.19.0.19)	3.75 (0.24.0.23)	2.62 (0.22,0.23)	2.21 (0.22,0.23)	1.99 (0.39.0.34)
Area	-2	-3.40 (0.07.0.07)	-3.16 (0.11.0.11)	-2.42 (0.09,0.09)	-2.13 (0.08,0.09)	-2.00 (0.23.0.20)
*y*1	0.7	0.00 (0.01.0.01)	0.12 (0.04.0.04)	0.49 (0.03,0.03)	0.63 (0.02,0.02)	0.70 (0.11.0.09)
Const	20	53.50 (0.41.0.42)	47.95 (1.91.1.94)	29.10 (1.19,1.20)	22.49 (0.67,0.69)	19.86 (5.26.4.43)

NOTES.
Models generating data:
$y_1 = 50 - \beta_{month} w + 4SES + 3sex - 2area + \varepsilon_1;\ y_2 = 20 + 0.7y_1 + 2SES + 2sex - 2area + \varepsilon_2$
$\sigma(\varepsilon_1) = 16$, $\sigma(\varepsilon_2) = 12$. Sample size=30000.
Values of explanatory variables*: month of birth* (1-12); *SES*(1-5); *sex*(0-1); *area*(1-4)
Average over 1000 replications' estimates.
In parenthesis: (st. dev of estimates over replications, mean se. of the estimates).
(1)β_{month}=-0.3; (2) β_{month}=-1; (3) β_{month}=-2

Results are summarized in Table A2. In the first column we report the true values employed for data generation. The second column refers to individual matching: the estimate of the coefficient of lagged performance is nearly 0 and the effects of the other explanatory variables are strongly overestimated. In the last column we report the results of imputed regression already shown in Table A1. The three middle columns refer to cell mean regression. We allow the coefficient of the instrument (month of birth) to increase from -0.3 (the estimated value in our empirical application) to -2. Bias diminishes as the coefficient increases, but it is still noticeable even with a large sample and high predictive power of the instrument.

In Table A3 we show the results of a battery of simulations aimed at evaluating the behavior of individual matching and cell mean regression with Fuller's correction for measurement error on lagged scores. This method requires an estimate of the reliability, i.e. the squared correlation between the observed explanatory variable (affected by measurement error) and its true counterpart. This quantity can be estimated by $\left(1 - \frac{\widehat{var}(\varepsilon_1)}{\widehat{var}(y_1)}\right)$.

Table A3. Comparison of alternative estimation strategies: individual matching and cell mean regression, with and without Fuller's correction

	True value	Individual matching (no correction)	Individual matching (Fuller correction)	Cell matching 600 cells (no correction)	Cell matching 600 cells (Fuller correction)
SES	2	4.79 (0.07,0.07)	4.27 (1.21,1.11)	4.33 (0.17,0.17)	1.85 (1.68,0.98)
Sex	2	4.08 (0.19,0.19)	3.71 (0.93,0.86)	3.75 (0.24,0.23)	1.90 (1.28,0.76)
Area	-2	-3.40 (0.07,0.07)	-3.14 (0.61,0.56)	-3.16 (0.11,0.11)	-1.93 (0.84,0.50)
*y*1	0.7	0.00 (0.01,0.01)	0.13 (0.30,0.28)	0.12 (0.04,0.04)	0.74 (0.41,0.24)
Const	20	53.50 (0.41,0.42)	47.34 (14.51,13.30)	47.95 (1.91,1.94)	18.25 (19.83,11.75)

NOTES.
Models generating data:
$y_1 = 50 - 0.3month + 4SES + 3sex - 2area + \varepsilon_1;\ y_2 = 20 + 0.7y_1 + 2SES + 2sex - 2area + \varepsilon_2$
$\sigma(\varepsilon_1) = 16$, $\sigma(\varepsilon_2) = 12$. Sample size=30000.
Values of explanatory variables*: month of birth* (1-12); *SES*(1-5); *sex*(0-1); *area*(1-4)
Average over 1000 replications' estimates.
In parenthesis: (st. dev of estimates over replications, mean se. of the estimates).
Fuller's method run with STATA, procedure Eivreg.

Despite measurement error does not meet CEV conditions, Fuller's method works well in terms of bias: average values of the estimates are close to real values. Standard errors, however, are considerably larger than those obtained with imputed regression.

Children repeating grades

This simulation has been carried out to evaluate imputed regression estimates with children repeating grades. We make children repeat 5th grade if their performance score is below a given threshold. The following year they move to 6th grade. The same rule applies to repetitions in grade 6. No children fail before grade 5.

Model (6) is estimated on regular children. As for dynamic model (7), we compare the behavior of the estimates when using only regular children and when including late children. We analyze two cases, with different shares of children failing the year. Since we introduced the issue of repeating grades when attempting to include immigrant children in the estimation, we add migrant status in the models. Given their poorer performance, low socioeconomic status and immigrant children are overrepresented among late children.

Results, shown in Table A4, confirm our theoretical expectations. We find no bias when analyzing only regular children, while we overestimate the effects when late children are included in the estimation.

Table A4. Imputed regression estimates with children repeating grades

		Only regular children		*Regular and late children*	
		Overall % repeating grades		Overall % repeating grades	
	TRUE VALUE	0% GRADE 5 10% GRADE 6	10% GRADE 5 20% GRADE 6	0% GRADE 5 10% GRADE 6	10% GRADE 5 20% GRADE 6
y_1	0.7	0.70 (0.11,0.09)	0.71 (0.12,0.10)	0.81 (0.11,0.09)	0.75 (0.11,0.09)
SES	2	2.00 (0.44,0.37)	1.97 (0.47,0.40)	2.33 (0.46,0.38)	2.30 (0.46,0.37)
2G	-3	-2.99 (0.51,0.44)	-2.97 (0.55,0.47)	-3.43 (0.54,0.44)	-3.26 (0.53,0.43)
Constant	10	9.93 (5.25,4.43)	9.63 (5.62,4.70)	0.53 (5.44,4.52)	3.31 (5.44,4.33)

NOTES.
Models generating data:
$y_1 = 50 - 0.3month + 4SES + 4mig + \varepsilon_1;\ y_2 = 10 + 0.7y_1 + 2SES + 2sex - 3mig + \varepsilon_2$
$\sigma(\varepsilon_1) = 16$, $\sigma(\varepsilon_2) = 12$. Sample size=30000.
Values of explanatory variables*: month of birth* (1-12); *migrant*(0,1) 20% migrants
Average over 1000 replications' estimates.
In parenthesis: (st. dev of estimates over replications, mean se. of the estimates).

Appendix B. Models with natives and second-generation immigrants

Table B1. Estimates of cross-sectional and pseudo-panel data models (natives and second-generation immigrants)

	READING			MATHEMATICS		
Variables	5th grade cross-section	6th grade cross-section	6th grade dynamic	5th grade cross-section	6th grade cross-section	6th grade dynamic
Costant	66.0***	58.8***	12.9*	61.7***	45.3***	-16.0**
Month[1]	-0.3***			-0.3***		
2G	-6.2***	-6.6***	-2.2*	-4.5***	-4.2***	
Female	0.7***	2.2***	1.7***	-3.2***	-3.7***	-0.5
Female 2G		2.9*	2.9*		2.6*	3.1**
Books[2]	2.3***	2.1***	0.4	2.1***	2.4***	0.2
ESCS	3.3***	3.9***	1.5***	3.0***	3.7***	0.5
North East	-0.7	-0.6	-0.1	-1.3*	0.9	2.4***
Centre	-2.6***	-1.3***	0.5	-2.1**	-2.8***	-0.6
South	-3.6***	-2.5***	0.2	-0.9	-5.1***	-4.1***
Islands	-6.3***	-6.5***	-1.8*	-4.0***	-8.4***	-4.1***
South-Isl 2G					3.6*	3.8*
y_1			0.718***			1.031***
R^2	0.135	0.161	0.163	0.094	0.153	0.157
sample size	27779	30882	30882	28542	30881	30881

*p_value<0.05; **p-value<0.01; ***p-value<0.005

NOTES. The estimation has been performed only on children born in 1999.

1. January=1,… December=12

2. 0=0-10 books; 1=11-25 books; 2=26-100 books; 3=101-200 books;4=>200 books

Cluster robust standard errors (Huber-White estimator, run with STATA). Clusters defined by classes.

Interactions effects are reported only when significant.