Re-Testing PISA Students One Year Later. On School Value Added Estimation Using OECD–PISA

Massimiliano Bratti*
Università degli Studi di Milano
and IZA, Bonn

Daniele Checchi#
Università degli Studi di Milano
and IZA, Bonn

In two provinces of Northern Italy the OECD PISA 2009 test was re-administered to the same individuals one year later. We show that: a) *cross-sectional school fixed effects estimates are volatile over time, given the high year-to-year attrition in the student population;* b) *longitudinal measures of school value added are more robust to student attrition, but require adequate control for sample selection of both schools and students;* c) *longitudinal measures provide still inadequate measures of teachers/schools contribution to student competences when the secondary school tracks students on ability, students can change tracks and/or drop out from schools.*
[JEL Classification: I21; J24].

Keywords: Italy; OECD-PISA; school assessment; student attrition.

* <*massimiliano.bratti@unimi.it*>, Department of Economics, Management and Quantitative Methods (DEMM), Università degli Studi di Milano.
# <*daniele.checchi@unimi.it*>, Department of Economics, Management and Quantitative Methods (DEMM), Università degli Studi di Milano.
We wish to thank for their collaboration: INVALSI (Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e di Formazione), in particular Piero Cipollone, Laura Palmerio and Sabrina Greco; IPRASE (Istituto Provinciale per la Ricerca e la Sperimentazione Educativa - Trento), in particular Francesco Rubino, and SREV (Struttura Regionale per la valutazione del sistema educativo - Valle d'Aosta), in particular Piero Floris and Paola Gallotta. We also thank Anna De Paoli for research assistantship in the field analysis conducted in Trento. Participants to the workshops and conferences "Improving Education through Accountability and Evaluation. Lessons from Around the World" (2013, Rome), the Fifth International Workshop on the Applied Economics of Education (2013, Catanzaro), the 25th Conference of the European Association of Labour Economists (2013, Turin), Enrico Rettore and Erich Battistin are gratefully acknowledged for their comments. The usual disclaimers apply.

1. - Introduction

Students and families are often interested in identifying which are the schools attaining the best performance, irrespective of whether this outcome is due to better students, better teachers, better resources or a combination of them. Researchers and policy makers are more interested in the contribution that schools and teachers provide to students' competences, sometimes indicated as *school value added* (VA, hereafter – see American Statistical Association, 2014). The availability of reliable measures of school VA raises schools' accountability, since policy makers can allocate resources in an efficient and effective way. The general consensus is that building a proper measure of school VA requires longitudinal data, which allow for measuring the increase in knowledge for the same individual overtime, some of which may be attributed to the attended school. However, recent contributions (Rothstein, 2009 and 2010) have shown that also longitudinal measures can provide biased measures if it is not possible to control for self-sorting of students

Our paper exploits data collected in an occurrence of PISA testing/retesting to explore the possibility of properly measuring the school VA in the framework of the Italian secondary school. This framework is interesting for at least two reasons: first, Italy has only recently started to build up a national system of school evaluation, and only very recently some analysis of school VA have begun to appear (INVALSI, 2016, chap. 7); second, the Italian secondary school is organised along three tracks (academic - *licei*, technical - *istituti tecnici*, and vocational - *scuole professionali*), with significant mobility across tracks and early school leaving (exceeding 20% at the time of the present survey). One may object that PISA survey is not intended to evaluate single schools, rather to measure the performance of national educational systems in a comparative perspective. However, PISA surveys measure "knowledge for life", which represents a concept of knowledge which is not necessarily curricular and which is assumed to be persistent over an individual's life cycle. This is probably closer than curricular competences to the economic concept of "human capital", which represents the *stock* of knowledge embodied in the individual and which contributes to increasing her productivity or quality of life (health, political participation, etc.),[1] and can be used to monitor regional differences or to promote educational reforms (like de-tracking – see Pons, 2011).

[1] Curricular competences may instead be year or age specific, *i.e.* they may be subjected to a very high rate of depreciation, and may not represent an optimal measure to evaluate the "social role" of schools.

Thanks to two field experiments conducted in two Northern Italian regions (Trento and Valle d'Aosta), PISA tests were re-administered in 2010 to the same students who were included in the 2009 official survey. To the best of our knowledge, this is the first paper to use data on a PISA re-test, and gives us the opportunity to investigate some important issues. First, the re-test exercise is useful to contrast cross-sectional measures of school VA, which are commonly available, with longitudinal measures. Second, the re-test was implemented on a voluntary basis in Trento while the entire student population was involved in Valle d'Aosta. This enables us to discuss the issue of panel attrition in two very different survey's contexts, and its relevance when the objective is to build longitudinal measures of school value added. Third, possibly as a consequence of different educational policies conducted at regional level, our analysis uncovers marked differences in the two "educational systems": the correlation between school fixed effects computed from cross-sectional data in 2009 and 2010 is high in Trento and low in Valle d'Aosta; in addition, when measuring school VA in a longitudinal perspective, the persistence in students' achievements is low in Trento and high in Valle d'Aosta. Fourth, we make an attempt at evaluating schools' contribution to a measure of skills (knowledge for life) which is closer to the concept of human capital than those measured by standardized curricular tests.

Our suggested interpretation of these two apparently contradictory results is rooted into the different selectivity of the two educational systems. Indeed, while in Trento only 8% of the students who were originally tested in 2009 dropped out or changed school in 2010, the percentage rises to about 21% in Valle d'Aosta. Through a simple economic framework in which an individual's school performance depends positively on the ability of her peers and negatively on the heterogeneity of the peer group, we suggest that higher "selectivity" (defined as a higher number of students dropping out or changing schools) is a possible determinant of both the lower correlation between cross-sectional measures of school VA and the higher year-to-year persistence in student test scores.

Our analysis also shows that, irrespective of different drop-out rates, longitudinal measures of school VA, based on panel data, are little sensitive to student attrition, *i.e.* to the fact that some students who participated in PISA 2009 for various reasons did not participate in the 2010 re-test exercise. However, if the students changed track from one year to the next (and this is more frequent at the end of grade 10, which is the modal grade attended by PISA students) the estimated VA is likely to be biased, since academic schools are losing weaker students, while vocational schools are gaining students from more selective schools.

Thus longitudinal measures in tracked school systems do not provide unbiased estimates of the true school VA due to the potentially different student selection and retention policies used by schools and educational systems.

The structure of the paper is as follows. The next section introduces institutional backgrounds of the two re-tests, whereas Section 3 illustrates similarities and differences between the conditions under which the two experiments were conducted. Our empirical strategy is illustrated in Section 4, providing the theoretical justification for the following empirical analysis; technical aspects related to selection bias and strategies to cope with it are reported in an appendix. The main empirical results are reported in Section 5 and discussed in Section 6. The final Section summarizes the main findings and concludes.

2. - Data and Context

Italy has always participated in the OECD-PISA project since its inception in 2000. Due to the lack of a national system of school assessment, the PISA survey became the first source of information on the performance of the Italian secondary school system, showing significant between-region and between-school variations (Bratti *et* al., 2007).[2] Table 1 shows that student performance varies along two dimensions: the type of track attended and the North-South latitude. Looking at country level, the average distance between an academic oriented track (*liceo*) and a vocational track (*istituto tecnico* or *scuola professionale*) is close to one and half standard deviation.[3] The geographic divide is as much impressive: other things constant, the North-Eastern part of the country obtains the highest average test score, closely followed by the North-West macro-region. The Centre and South-Islands then follow, with a gap well above half standard deviation.

In this paper we will focus onto two Northern regions which have conducted two re-tests of students for research purposes: Trento and Valle d'Aosta (see Figure 1). The Autonomous Province of Trento is a small province of half million inhabitants, located in the North-East part of Italy, close to the Austrian border. Valle d'Aosta is an even smaller province (more precisely a region which contains only one province) in terms of population (128,000 inhabitants), located in the North-West of the country, for centuries under the rule of the French-origin royal

2 Many regional governments financed the oversampling of the PISA survey in order to obtain adequate information to assess local school systems.

3 In the international sample PISA scores have a mean of 500 and a standard deviation of 100.

family Savoy, which one century and half ago succeeded in unifying the Italian nation. As other bordering regions (like Friuli Venezia Giulia), due to political reasons related to the difficult process of country unification both provinces enjoy greater autonomy in administration (like school design) and revenue collection (not participating in the cross-region redistribution). Nevertheless both follow the Italian scheme of a tracked secondary school system, even if their regional-based vocational tracks enjoy (at least in the Trento region) better standards, as in the German tradition. When looking at student achievements through the PISA lens (see again Table 1) we observe that students from Trento's or Valle d'Aosta's schools obtain results that are in line with the bordering macro-regions, and better than the Central and Southern regions. This is mostly attributable to the relative performance of state vocational schools, which score almost half of a standard deviation above the schools of the same track in the rest of the country. In addition to macro-inequalities, and despite the existence of tracks which attract different types of students, there is also significant between-school variation.

Before entering the analysis of school quality, the differences across regions and across school tracks raise questions about student allocation across tracks. When we look at the variance decomposition (Table 2), we notice that the between-track variance is higher in Trento compared to the rest of the country, while it is lower in the case of Valle d'Aosta. By contrast, there are no significant differences across areas when considering the between-school variance. This suggests that the relevant choice in Trento is the school track (since school quality seems rather homogenous within tracks), while in Valle d'Aosta there is an additional problem of choosing both the track and the school. The literature points to family background as the main factor driving students into different tracks. Without resorting to multivariate analysis, simple descriptive statistics (Table 3) suggest that sorting by social background may be less pronounced in Trento *vis-à-vis* the rest of the country.[4] While in the rest of Italy students from better backgrounds (higher social prestige associated to parental occupation, better parental education as measured by years of education and better ESCS score)[5] are gathered by the academic track (*liceo*), then by technical schools and eventually by voca-

[4] A more rigorous statistical analysis (ordered probit model) does not identify a precise pattern of sorting across regions, probably due to the lack of a proper measure of student ability.

[5] ECSC stands for index of Economic, Social and Cultural status and is derived from the highest occupational status of parents, highest educational level of parents (in years of education), family wealth, cultural possessions and home educational resources. It is recoded as zero mean variable.

tional schools, in Trento this selection process is less pronounced, and students in vocational schools seem better endowed with parental resources (at least *vis-à-vis* students in technical schools). What is common to the whole country is that regional vocational schools (which are characterised by shorter duration, three instead of five years) attract students from poorer backgrounds. Despite Trento and Valle d'Aosta being hardly representative of the entire country, they do represent two interesting case studies, which are sufficiently dissimilar one to the other to highlight different problems connected to the measurement of school VA using longitudinal data.

3. - Description of the PISA Re-Tests

In the year 2009 a decision to re-test PISA students was independently made by two local educational authorities, following the advice of the local supervisory committees. In the case of Trento the opportunity was given by an effort at data collection, needed for the drafting of a biannual report on the state of Trento's schooling system. In the case of Valle d'Aosta the occasion came from the debate on the benefits to student learning deriving from bilingual education (Italian and French). As easily conceivable, the two projects underwent different negotiations with local schools' headmasters, the final outcome being different strategies of implementation which do not grant full comparability of the two exercises. The main differences concern student participation (sampled in Trento, universal in Valle d'Aosta), school participation (voluntary in Trento, mandatory in Valle d'Aosta), test score reweighting (performed by INVALSI, the Italian agency of school assessment, in Trento and by ACER[6]-PISA consortium in the case of Valle d'Aosta) and the language used for the test (only Italian for Trento, Italian or French for Valle d'Aosta). Nevertheless, the testing strategy is identical, and this grants sufficient comparability of the estimates across the two exercises. The potential sample selectivity issues related to the structure of two re-test exercises are discussed in the Appendix: our overall conclusion is that selection biases are limited in the Trento experiment (because negative selection of schools somehow balances positive selection among students), while there are biases in the Valle d'Aosta experiment, due to positive selection of students associated with early school leaving or track change.

[6] ACER stands for Australian Council for Educational Research.

3.1 *Structure of the PISA Test in 2009 and 2010*

Following discussions with the PISA's headquarter and with the Italian national agency for school assessment (INVALSI), it was decided to resubmit in 2010 the same PISA booklets already used in 2009 for two main reasons. First, questions are not related to academic curricula set by the Ministry of Education, but they are intended to measure «… students' ability to complete tasks relating to real life, tapping a broad understanding of key concepts, rather than limiting the assessment to subject-specific knowledge» (OECD, 2010, p. 24). As a consequence, administering the test a second time does not prevent checking for improvements in pupils' literacy levels. Second, the test is available in 13 different versions of an equivalent level of difficulty (booklets) and it has, therefore, been possible for each participant to minimize the number of identical questions between the two test sessions. In fact, each student has been assigned a booklet in 2009 and one in 2010, according to the scheme described in Table 4.[7] The students then had to answer several questions, three quarters different and one quarter identical between sessions; the test is focused on the assessment of reading skills, since the main PISA test's focus in 2009 (and therefore also the 2010 retest) was on students' comprehension of written texts, although half of the booklet concerns competences in science and mathematics. In both events the student, the school and the parent questionnaires were not administered twice, on the presumption that relevant information would have not changed after one year.

3.2 *Participating Schools and Test Administration in Trento*

During the winter 2009 all 49 secondary schools in the Trento province whose pupils had taken the test PISA 2009 were contacted and asked to resubmit an equivalent test to the same students, and only 35 agreed to a second administration of the test.[8] We discuss in the Appendix the potential selection bias arising from

[7] Each booklet consists of four sections and each section is identified by a type (M = mathematics, S = sciences and R = reading) and an index. There are, therefore, three different sections for mathematics (M1, M2 and M3) and sciences (S1, S2, S3) and seven different sections for reading (R1, ..., R7). It has to be noted that the allocation of the booklets was such that each student answered in 2010 some questions that had already been responding in 2009. This overlap occurs for all students, but only for a quarter of the questions, which is a single section of the test (grey cells in Table 4 show the overlapping section). For example, all students who had received the booklet 1 in 2009 were assigned the booklet 7 in 2010, where section M3 is common to both years.

[8] As our main focus is on VA estimation for upper secondary schools, we dropped from the analysis one lower secondary school which was sampled in PISA 2009.

schools' and/or students' non-participation to the 2010 re-test. Whenever possible, for each school the reference person who had been responsible for testing in 2009 was contacted again in 2010. After meeting all reference persons to illustrate the potential test administration problems[9] (April 16th, 2010), in May 2010, the booklets and the lists containing the identification codes of the students were transmitted from INVALSI to the local agency *Istituto Provinciale per la Ricerca e la Sperimentazione Educativa* (IPRASE). Schools were given a window of two weeks for administering the test. The schools were required to return to IPRASE the booklets compiled within 48 hours from completion of the retest. In June, the tests were sent back from IPRASE to INVALSI which carried out the correction and scoring. INVALSI was constantly present in each stage of the re-test exercise, providing the technical assistance for test administration, correction of open questions and estimation of student scores (including the plausible values).[10]

3.3 *Participating Schools and Test Administration in Valle d'Aosta*

Following a request from the local educational authority for evaluating the impact of bilingual education (Italian and French) onto student learning, an agreement was signed with the OECD-PISA consortium in order to administer the French version of the questionnaire to a random sample of students from Valle d'Aosta's secondary schools. In order not to alter the standard national assessment conducted in 2009, it was decided to re-test Valle d'Aosta's students in 2010 using an identical scheme of booklet rotation to that just described (Table 4). Given the small size of each age cohort, all 15-year-old students enrolled in regional upper secondary schools (including private ones) were tested in 2009 (universal coverage). The very same students who were still enrolled in one of the

[9] In the meeting some doubts emerged about absent pupils. In the case of pupils who were absent on the day of the retest, they were not given the opportunity to be tested in a second session. However, schools were asked to report the reasons of the student's absence from school: transfer to another school, authorisation denied by parents, school drop-out, or "standard" truancy. Schools were instead requested to administer the tests to the students who were sampled for PISA 2009 but were absent during the test day, and who were instead present in 2010. Teachers in charge of the test administration were required to try to replicate in 2010 the same testing conditions existing in 2009 (duration, rules of conduct, rooms' characteristics, time of the day). The main goal was to make the two testing exercises as similar as possible in order to minimise the incidence of framing problems.

[10] Starting from original students' answers to the tests, INVALSI estimated the plausible values in the Trento 2009 sample provided by ACER-OECD, and then used the same algorithm to impute the 2010 plausible values. Technical details on the estimation procedure are reported in DI CHIACCHIO C. *et* AL. (2010). Thus scores in 2009 and 2010 are comparable.

regional secondary schools were retested on the same day (April 13th, 2010) one year later, but half of them (randomly determined at school level) obtained a booklet in Italian and the other half got it in French (the language spoken in the bordering region, even if the local variant – *patois valdôtain* – has dialectal inflexions).[11] In this way it became possible to test both the potential increase in competences (comparing test scores in Italian over the two years) as well as the advantage/disadvantage of using Italian or French while tested. The tests were then mailed directly to ACER which returned the plausible values computed according to the international scale.

The universal coverage of the 2009 survey highlights a different aspect of attrition in longitudinal studies. Looking at the numbers reported in Table 5, we observe that 879 students took the test in 2009, but only 736 were traced in the following year, losing 16% of the initial student body.[12] Among the 143 lost students, 75 are officially reported as absent/truant the day of the 2010 test, 16 are unmatched among the two samples and the remaining 52 are truly drop-outs (since compulsory education ends at the age of 15 in Italy) or movers outside the region. Within the 736 students included in the longitudinal component, only 696 persisted in the same track, while 40 (equivalent to 4.5% of the initial student body) changed school track within the local schooling system.[13] In order to evaluate the school VA, we have to rely on this permanent component only, but we try to account for the potential bias due to these different sources of attrition.

4. - Econometric Models of School Value Estimation

Various definitions of school VA exist in the academic literature. In general, school VA can be defined as the contribution that schools give to students' competences over and above contextual factors. A good school VA model should take into account the characteristics of the student intake, which are likely to affect

[11] Details can be found in Assessorato Istruzione e Cultura della Regione autonoma Valle d'Aosta Dipartimento Sovraintendenza agli Studi - Ufficio Supporto Autonomia Scolastica (2012), *La Maîtrise de langue française. Rapport Régional PISA 2010* - Edition pour la Vallée d'Aoste.

[12] ACER-OECD released a file containing 752 observations appearing in both surveys, and we matched them to the public file of PISA 2009 survey using the test results. We were able to match 736 students, thus losing 16 additional students which are reported in Table 5 as outside the local schooling system.

[13] Table 5 shows only track changes (40 students), but comparing the school code over the two years allows for the identification of 12 additional students who changed school within the same track (and which are consequently excluded from the panel component).

students' competences irrespective of schools they are enrolled in. Educational VA can be evaluated at different levels of aggregation. One can be interested in the school's VA or in teachers' VA, as in Rothstein (2009). In this latter case in order to distinguish the effect of teachers from that of the peer group it is necessary to have information on different classes within the same school, which is not the case of OECD-PISA, where students are sampled out of schools and are not representative of specific classes.

If we intend to use the above field experiments to evaluate schools we need to make two important assumptions. The first one is that PISA tests, measuring "knowledge for life" and not curricular competences, can be an adequate outcome to evaluate schools' performances. The second assumption is that these competences for life, which are also correlated to idiosyncratic factors (unobservable talent, family background, contextual factors), can be increased via effective schooling. In the light of policy interventions intended to enhance low-performing educational systems which the OECD-PISA surveys have provoked in many countries, we view these two assumptions as consistent with PISA project.

An important reference for researchers aiming at estimating educational production functions (EPF) is Todd and Wolpin (2003) who discuss the problems posed by this difficult object. The authors carefully describe all the limitations of cross-sectional measures of school VA, and the extreme richness of data needed to obtain unbiased estimates of the effects of inputs entering the EPF, which then allow for an unbiased calculation of school VA. Unfortunately, in most cases researchers have to work with much less rich data, and this is also our case. Todd and Wolpin (2003) highlight how the availability of longitudinal data represents a first step towards a better understanding of the process of knowledge production. However, Rothstein (2010) shows that the absence of random allocation of students to classes/teachers/schools prevents an unbiased estimation even of the longitudinal VA of schools, due to serial correlation of students' test scores.

In order to illustrate these problems, we assume that the data generating process for student literacy (T_{ijt}) can be expressed through the following EPF $f(.)$:

$$T_{ijt} = f\left(B_{it}, VA_{jt}, a_i, \varepsilon_{ijt}\right) \tag{1}$$

where i, j and t are subscripts for individual, school and time respectively.[14] T_{ijt} is the test score of student i in school j surveyed in year t, while B_{it} are current family inputs. Time-invariant unobservable individual components (like innate ability, self-confidence) are represented by a_i. The contribution of school j (including the effect of teachers' quality, motivation, school finances, etc.) to student i's performance at time t is defined as the school valued added VA_{jt}. ε_{ijt} captures the effect of time variant unobservable variables.

As it is often customary in the EPF literature, we adopt a linear specification. For illustrative purposes, let us assume that the student performance equation is given by the following EPF:

$$T_{ijt} = \alpha_0 + \alpha_1 \cdot t + \alpha_2 B_{it} + \lambda_j + \tau_j \cdot t + \left(a_i + \varepsilon_{ijt}\right) \tag{2}$$

The contribution of school j to student i's performance at time t is captured by a fixed effect (λ_j) and by the interaction $\tau_j \cdot t$ (a school-specific age trend); ε_{ijt} is a white noise. Equation *(2)* shows that two potential measures of VA are available. In cross-sectional studies, the researcher will be able to estimate only λ_j, while in longitudinal studies, *i.e.* when repeated observations on test scores are available, both the school fixed effect and the school-specific age trend could in principle be estimated. The specification also shows that in the absence of a measure of individual ability it is difficult to interpret λ_j as school value added, since these coefficients also capture the effect of other time-invariant schools' attributes (including the peer effect). By contrast, $\tau_j \cdot t$ can be considered as a more correct measure of school VA as it refers to the average improvement overtime in the literacy levels of the students enrolled in school j.

Estimation of equation *(2)* with ordinary least squares (OLS) poses several problems. Indeed, individuals are unlikely to randomly self-sort across schools, causing a correlation between individual ability, which is unobservable and enters

[14] Here we do not follow a "cumulative approach" for knowledge production, that is we do not let past literacy enter the current production of literacy (thus equation *(1)* does not correspond to models VAM1-VAM2-VAM3 proposed in ROTHSTEIN, 2010) So doing we focus on the best-case scenario in which the availability of two adjacent time observations (T_{ijt} and T_{ijt-1}) of test scores (as in our case) can help the researcher to estimate the "true" EPF. This would not be possible if EPF had a cumulative form since T_{ijt-2} (which is not observed), for instance, enters T_{ijt-1}. However, we will see in what follows that also in our specification past performance will enter current performance through the role of unobserved individual ability (see equation *(5)*).

the error term, and school fixed effects and age-specific school trends. The estimates of all coefficients suffer from an omitted variables bias.

This can be partially addressed having repeated observations on student performance. Lagging equation *(2)* one period we obtain

$$T_{ijt-1} = \alpha_0 + \alpha_1 \cdot (t-1) + \alpha_2 B_{it-1} + \tau_j \cdot (t-1) + \lambda_j + \left(a_i + \varepsilon_{ijt-1}\right) \tag{3}$$

from which we can easily derive an expression for individual ability

$$a_i = T_{ijt-1} - \alpha_0 - \alpha_1 \cdot (t-1) - \alpha_2 B_{it-1} - \tau_j \cdot (t-1) - \lambda_j - \varepsilon_{ijt-1} \tag{4}$$

and replacing *(4)* in *(2)* we eventually get

$$T_{ijt} = T_{ijt-1} + \alpha_1 + \alpha_2 \cdot \left(B_{it} - B_{it-1}\right) + \tau_j + \left(\varepsilon_{ijt} - \varepsilon_{ijt-1}\right). \tag{5}$$

Equation *(5)* clearly shows that a measure of past literacy T_{ijt-1} acts as a sufficient statistics for individual unobserved ability. In this case, under the assumption that $\Delta\varepsilon \equiv (\varepsilon_{ijt} - \varepsilon_{ijt-1})$ is uncorrelated with school VA, consistent estimates of schools' VA can be retrieved from the OLS estimates of the school fixed effects τ_j. In addition, equation *(5)* also allows for consistent estimation of the effect of family inputs α_2 if the variation in family inputs $\Delta B \equiv (B_{it} - B_{it-1})$ is uncorrelated with $\Delta\varepsilon$.

Two things are noteworthy. First, past performance enters equation *(5)* with a unitary coefficient. However, in the data we do not observe true literacy but an imperfect measure of it, namely $T_{ijt} = T^*_{ijt} + m_T$ where m_T is the measurement error in the true literacy score T^*. Classical measurement error (e.g., m_T could be considered as pure luck) in the absence of other regressors in *(5)* could lead to an attenuation bias, and the estimated coefficient on T_{ijt-1} could be less than one. With other covariates, the sign and magnitude of the bias depends on the correlation between T_{ijt-1} and the other explanatory variables. Second, if family questionnaires are not re-administered every year (as in our case for the Trento's and for Valle d'Aosta's re-tests), we only have one measure of family background (B_{it-1}). When the two observations of test scores are close enough in time, the assumption that $\Delta B = 0$, *i.e.* of no variation in family background, can be considered quite innocuous. In this case the EPF becomes

$$T_{ijt} = T_{ijt-1} + \alpha_1 + \tau_j + \left(\varepsilon_{ijt} - \varepsilon_{ijt-1}\right) \tag{6}$$

which corresponds to VAM1 model in Rothstein (2010). However, if $\Delta B = 0$, including only past family inputs leads to the specification

$$T_{ijt} = T_{ijt-1} + \alpha_1 - \alpha_2 B_{it-1} + \tau_j + \left[\alpha_2 B_{it} + \left(\varepsilon_{ijt} - \varepsilon_{ijt-1}\right)\right] \tag{7}$$

where the sum in brackets represents the new error term. Then it is clear that the estimates of the contribution of past family background are biased if $\alpha_2 \neq 0$ and family background is correlated over time (as it is likely to be the case). As T_{ijt-1} includes past family background, also the coefficient on past performance is biased (under the same hypothesis of serial correlation). The estimates of the school fixed effects are unbiased only under the assumption of no school selection according to *current* family background.[15] This additional assumption is not too strong if the choice of the specific school has been made in the past, as it happens in the case of Italy where upper secondary schooling starts at age 14, according to persistent family characteristics which are well proxied by B_{it-1}. However, as discussed in the Appendix, when changes of school occur within the sample period (as in the case of Valle d'Aosta for a significant share of students – see also Table 5) VA estimates must be taken with caution, because they are likely to be biased.[16]

[15] The inclusion of past performance in *(7)* will partly account for this selection if family background is serially correlated. A comparison between the estimates of *(6)* and *(7)* can also suggest whether the assumption $\Delta B = 0$ is credible, as in this case equations *(6)* and *(7)* should give very close estimates of the school fixed effects τ_j.

[16] The direction of the bias is in principle undetermined, unless one is able to model the process of school/track change of the students. Table 5 suggests that most of the school mobility is a downward mobility (*i.e.* towards less demanding tracks, as vocational tracks in Valle d'Aosta are characterised by lower performance, not so in the case of Trento – see Table 1). If this is the prevailing flow, we expect a positive bias in the estimated VA for Valle d'Aosta; see the discussion in the Appendix.

In this context, sudden changes in family background, unless very extreme (e.g., parental death), are rather unlikely to push individuals to change school or drop out. Summarizing, under equation *(7)* being the true data generating process for educational performance, we can get unbiased estimates for the school VA's as long as we can exclude: *a)* measurement errors in measures of students' past performance; *b)* variations in current family background; and *c)* student self-selection into the different schools according to unobservable characteristics. Owing to the characteristics of our data, these will be our maintained assumptions throughout the rest of the paper.

5. - School Value Added Estimation

In this section, we report the estimates of school VAs obtained from EPF estimation for Trento and Valle d'Aosta. Given the limited number of students per school, we are forced to assume that the contribution given by school *j* to its students' literacy levels (the VA) is the *same for all students*, *i.e.* the school has an "intercept effect" only.[17] Table 7 shows the results for Trento. All regressions are estimated on the panel component so as to isolate the effect of using different specifications of the EPF from potential differences in the composition of the samples.[18] In column (1) we have reported a simple specification which does not include school fixed effects (SFE, hereafter), using the test scores in 2009 as the dependent variable. The age and grade of the student turn out to be positively associated with performance in the PISA test. The same happens for cultural capital (the number of books at home), and the family socio-economic index (HISEI). Having attended kindergarten and being a first-generation immigrant are marginally significant, with a positive and a negative sign, respectively. In column (2) we include SFE: the inclusion of SFE generally has the effect of reducing the coefficients' magnitude of the significant regressors, suggesting that self-selection of schools and/or students is taking place (see the Appendix). Cultural capital, HISEI and first-generation immigrants all reduce in size. Column (3) es-

[17] This assumption could be relaxed by including interaction terms between the school indicators and student characteristics but this is feasible only when a large number of students are sampled for each school.

[18] We should keep in mind that unless we adopt random effect models, we are forced to choose one excluded school, which then represents the benchmark case against which we compare the relative effectiveness of the remaining schools. In the present analysis, we have left out the final school (that with the highest code), which in both regions is a regional vocational school.

timates the same model of column (1) but using the 2010 test score as the dependent variable. The results in column (1) and (3) are qualitatively and quantitatively very similar. The same is observed when comparing column (2) and column (4), which both include SFE. The coefficients on grade, cultural capital and first-generation migrants tend to be slightly lower in 2010.

Columns (1) to (4) estimate performance in levels, while column (5) specifies a "value added model", including the test score in 2009 as an additional regressor for the test score in 2010. Notably, the only significant regressors turn out to be past test performance and first generation immigrant status. The coefficient on the past test score is 0.304 (column (7) controlling for self-selection), well below the coefficient of one predicted from the model outlined in Section 4. As we mentioned earlier, this may be partly due to measurement error and the fact that test scores are likely to contain some noise, and not to perfectly measure true literacy. In column (6) we have included only past performance among the regressors. Under the assumption that changes in student characteristics in two adjacent years are almost null, we should expect very similar estimates of the SFE from the models (5) and (6). Indeed, the coefficients on past scores in the two columns are not statistically different.

Below Table 7 we have reported the correlation and the rank correlation between the estimates of the SFE obtained with various models. We focus on the rank correlation because ranking schools is often in the interest of educational policy makers. First, when using specifications in levels, the rankings of SFE in 2009 and 2010 are highly correlated (0.88). When switching to a longitudinal value added model, the rank correlation is much more similar with respect to the specification in levels in 2010 than in 2009 (0.81 *vs.* 0.98), but is in both cases very high. All in all, these estimates suggest that if one focuses on the panel component of the dataset, that is on the students which participated in the two testing exercises, and there are no significant alterations of the student body attributable to drop-out or school changes, the specific model used to estimate SFE does not make a huge difference for the estimation of schools' VA. However, estimates of the school VA may be influenced by the student self-selection in the 2010 retest. For this reason in column (7) we have reported the results of the estimates controlling for the propensity score, *i.e.* the probability of having participated in both tests,[19] and also the rank correlation with the SFE estimated with such model. As shown by column (7) of Table 7 the propensity score (PS, hereafter)

[19] For the specification of the PS we used the model in column (3) of Table A1.

is not statistically significant, suggesting that the inclusion of past performance is sufficient to control for the potential self-selection of students according to past (or prospective) performance. Consistently, the rank correlation of the SFE of model (5) and (7) is almost one. Our analysis suggests therefore that in the absence of track changes or selective drop-out (like it was the case for Trento) controlling for past performance is likely to address all ability-related potential estimation bias generated by panel attrition.[20]

When we repeat the exercise in the case of Valle d'Aosta (see Table 8) we find a similar attenuation of the coefficients on individual characteristics when SFE are included (columns (2) and (4)), confirming that students are (at least partially) sorted in schools according to individual characteristics. When we use past performance as a control (column (5)) we observe a much higher first order autoregressive, *i.e.* AR(1), coefficient than what we obtain in the case of Trento (0.63 *vs.* 0.30)[21] while many individual characteristics still retain statistical significance (gender, modal grade, parental education, availability of books and immigration status). Controlling for self-selection into the panel sample (column (7)) does not change the results much, also given the absence of reliable exclusion restrictions. Eventually, when we restrict to the reduced form represented by equation *(6)* (see column (6)), while controlling for the test language, we find that the AR(1) coefficient is still significantly different from one (test F = 45.11 (0.00)) but much higher in magnitude than for Trento. When we move to the rank correlation among the estimated measures of school VAs, we observe that single year cross-sectional measures are not even correlated; the correlation increases when we con-

[20] Yet we are unable to fully account for unobservable components related to students' ability in measuring schools' contribution to student test scores. If we net out these components by taking first differences in test scores as our dependent variable (model VAM1 in ROTHSTEIN J., 2010) and we estimate school fixed effects, we find limited correlation with previous measures of VAs (see the final row of the correlation matrix in Table 7). Visual inspection of this alternative measures of SFE suggests that in the case of Trento school rankings are rather different when we consider single-year measures obtained from cross-sectional data, but when we use past performance as an additional control the problem of self-selection is minimised. However, if can get rid of student time-invariant unobservable characteristics through the use of first differences in students' scores, school rankings appear quite different. But the use of individual first differences imposes the restriction of a unitary coefficient for T_{ijt-1}, which is clearly rejected in our data (see Table 7). This explains the low correlation between the SFE computed under this strategy and the ones previousely estimated.

[21] The difference between the two provinces disappears when we replace individual data with school averages: 0.99 (s.e. 0.13) for Valle d'Aosta against 0.93 (s.e. 0.06) for Trento, but these estimates are computed only over 22 and 35 observations, respectively.

sider longitudinal measures, but it does not reach the high value obtained in the case of Trento.[22]

6. - A Suggested Interpretation

In this section, we propose a potential reading of our results, and in particular of the differences found between the estimates using the two re-tests. The main differences are summarised in Table 9. First, the correlation between the cross-sectional measures of school VAs in 2009 and 2010 is very high in Trento (0.92) while it is almost zero in Valle d'Aosta (-0.02). Second, the coefficient on past test score is small in Trento (ranging between 0.26 and 0.30 - see Table 7) and much larger in Valle d'Aosta (between 0.63 and 0.78 - see Table 8).

In what follows we propose a possible interpretation of these two empirical facts based on *differential school selectivity in the two provinces*. Here, selectivity must be interpreted as *dynamic selectivity*, which is the change in the student body recorded over time in a specific school. Schools may be rather effective in inducing students to switch schools or drop-out by means of bad marks, retention and di-sciplinary measures; but they may also be quite effective in contrasting these changes in order to retain students, using remedial courses, guidance and/or financial support. We define selective schools those recording a high number of school switchers or drop-outs.[23] We intend to show that different degrees of selectivity may produce the contrasting evidence recorded in the two regions (highlighted in Table 9).

Let us suppose that student competence in school j at time t (the PISA test score in our case) is an increasing function of past year's performance, individual ability, peer group's average ability and the degree of homogeneity in (abilities of) the peer group:

$$T_{ijt} = T_{ijt-1} + a_i + \underbrace{\alpha \cdot \bar{a}_{jt} - \beta\sigma^2_{jt}}_{SFE_{jt}} + \varepsilon_{ijt} \qquad (8)$$

[22] What is more surprising is the higher correlation between the rankings obtained with and without controlling for individual student fixed effects in Valle d'Aosta (last row of the correlation matrix in Table 8): while in the case of Trento the rank correlation between longitudinal SFE measures with and without controls for unobservables was low, in the case of Valle d'Aosta it exceeds 0.7, suggesting that in the latter case these student components do not play any role.

[23] This is different from high selectivy at entry in terms of student ability, which may produce the opposite result (*i.e.* low drop-out and school changes).

where T_{ijt} stands for the PISA score of individual i in school j at time t, a_i is her level of (unobservable) ability (time-invariant by assumption), $\bar{a}_{jt} = a_{-it}$ is the average ability of her peers, σ^2_{jt} is the peer group's variance in ability and ε_{ijt} an idyonsincratic error terms. Individual performance is assumed to depend positively on an individual's past performance and ability, on the average level of ability of her peers, while it is negatively affected by school heterogeneity. The latter can be motivated by the difficulties of teaching to individuals with different levels of ability (*i.e.*, a "teaching quality" effect) or by class disruptive behaviour *à la* Lazear (2001).[24] Due to non-observability of abilities, the third and fourth addends in equation *(8)* jointly determine the SFE for school j at time t. In our analysis we are not able to disentangle the separate effects of the peer group and peer heterogeneity, which are both subsumed by the SFE. Given this simple setting, we may wonder what would be the effect of school VA estimation of having two educational systems (of the two different provinces) with different *degrees of selectivity*.[25] A high *dynamic selectivity* of the province's school system means that as time goes by, the school intake in terms of peer group's quality will tend to change. In equation *(8)* we face two effects. First, in the high-selectivity system the estimates of school VA in two adjacent years, t and $t + 1$, are likely to be less correlated than in the low-selectivity system, where average ability and its variance tend to be more persistent overtime. This effect can be particularly sizable in the first years of upper secondary schooling that also are the most selective, and in which the two PISA tests were administered (second and third years of upper secondary education, corresponding to grade 10 and 11, respectively). This is consistent with the first empirical fact that cross-sectional measures of school VAs show a higher correlation in Trento than in Valle d'Aosta. This however also poses some methodological issues, as cross-sectional measures of VAs for the same

---

[24] An additional justification for the inclusion of the variance among peers with a negative effect can be obtained by the existence of strong complementarities in individual abilities in the EPF (BENABOU R., 1996*a* and 1996*b*). In the limiting case where the elasticity of substitution goes to zero, the educational production function takes the form

$$T_{ijt} = \left[a_i^{\sigma} + (n-1)\, a_{-i}^{\sigma}\right]^{\frac{1}{\sigma}} \xrightarrow[\sigma\to\infty]{} \min\left[a_i, a_{-i}\right]$$

and the individual performance happens to be constrained by the lowest of peers' ability.

[25] We have shown in the Appendix that in the case of Trento the rate of student (panel) attrition in the schools which did participate in the 2010 re-test is around 20% (18.8%), but only about 8% are school drop-outs or school switchers. In the case of Valle d'Aosta, the incidence of school drop-out and school switchers is more than double (22% - see Table 5).

school can be very volatile from year to year, and moreover do not "penalize" schools for the potentially high number of drop-outs or school switchers. Actually, the SFE in *(8)* could increase overtime if a school exercises cream-skimming among its student body (corresponding to an overestimate of VA due to student sorting-out of the schools).

Let us recall that according to equation *(8)* individual past performance T_{ijt-1} can be expressed as

(9)
$$T_{ijt-1} = f\left(a_i, \bar{a}_{jt-1}, \sigma^2_{jt-1}\right).$$

Then in less selective school environments both the average level of peer group's ability and its variability will be highly correlated over-time. This means that in those environments lagged performance will be more correlated with the current school VA (the SFE), resulting in a lower AR(1) coefficient in the econometric specifications that controls for SFE. This is consistent with the evidence of a lower AR(1) coefficient in Trento than in Valle d'Aosta, *i.e.*, the second empirical fact that we observed.

The same argument could also explain why in the case of Trento we do not find any statistically significant association between family background characteristics and student performance after controlling for SFE (see column (5) in Table 7), while some student characteristics turn out significantly affecting the performance in the case of Valle d'Aosta.

A possible reason for the high selectivity of the school system in Valle d'Aosta is that students are initially more mismatched to schools, e.g., they often chose schools which are not aligned with their aspirations and levels of ability. This may stem, for instance, from better school guidance in Trento than in Valle d'Aosta. Better matches between students and schools imply similar family background characteristics across students, *i.e.* a more homogeneous school intake that is captured by the SFE. In schools where students are more mismatched and have more heterogeneous background characteristics, the SFE will not be a good proxy of peer characteristics, some of which may turn out to be significantly associated with school performance. Indeed, in the case of Trento we observe that individual past performance captures almost all relevant information at the individual level (see columns (5) or (7) in Table 7), and SFE are highly correlated across years (rank correlation of cross-sectional estimates is 0.88 – see bottom line of the correlation matrix associated to Table 7). A quite different situation emerges in the Valle d'Aosta's re-test. In this case including past performance does not eliminate the statistical significance of other individual characteristics (see column (5) in Table 8 – these effects are attenuated when we account for potential self-selection into the sample, as done in column (7) of the same table) and SFE are less correlated across survey years (correlation is null or even negative).

Despite the limited degrees of freedom, we can estimate an AR(1) coefficient for the panel component of students in each school. By restricting the set of regressors to gender, age, grade attended and a proxy for family background (number of books available at home), we estimated an AR(1) coefficient for each school, and plotted it against a measure of school (social) heterogeneity, namely the standard deviation of the prestige associated with parental occupations of the enrolled students. These scatter-plots are shown in Figures 2 and 3 which suggest a greater variation for the AR(1) coefficients in the case of Trento (Figure 2) compared to the case of Valle d'Aosta (Figure 3). In both regions we find a weak but positive correlation between social heterogeneity and persistence in individual test scores: when the social environment is more homogeneous, past performance in learning has a lower correlation with current performance, while on the contrary it increases in more heterogeneous environments.

Our suggested interpretation finds additional support in Table 10, which highlights the differences between cross-sectional and longitudinal estimates of gradients of individual family backgrounds and school tracks. In the table only some coefficients are shown, but the estimated models are fully equivalent to those presented in Tables 7 and 8 (except the fact that SFE are replaced by school track fixed effects). In columns (1), (2) and (3) we present the results referring to the panel component of the Trento re-test: the first two columns refer to cross-sectional estimates, while the third one exploits the longitudinal dimension by including past performance as an additional regressor. Columns (4), (5) and (6) replicate the same exercise for Valle d'Aosta; in order to account for possible distortions induced by different test languages, columns (7) and (8) restrict the analysis to the subsample of students who took the test in Italian. We notice that the introduction of past test performance as an additional regressor generally reduces the correlation between a student's literacy level and her characteristics, but this reduction is more pronounced in the case of Valle d'Aosta, in which schools have a more heterogeneous student body. With our data we are unable to ascertain whether curricular differences, teaching quality and/or contextual effects (e.g., peer groups) drive these effects, as well as whether there is any role for individual effort. More detailed information would indeed be needed to discriminate further among these alternative explanations.

7. - Concluding Remarks

What can be learned from the two PISA re-tests described in this paper? First of all, we have shown that cross-sectional measures of school value added (*i.e.* those obtained by educational production functions not controlling for past test scores) face remarkable problems of non-random attrition. In educational settings characterized by high student attrition, this could lead to very volatile measures of VAs which are difficult to interpret by both the public and policy makers. In the case of Valle d'Aosta, for instance, we have shown that the correlation between the estimated school VA in 2009 and 2010 is close to zero.

We have then contrasted the cross-sectional measures with longitudinal measures of school VAs. Here we face two main issues: first, in settings characterized by low student attrition (drop-out or school switchers), longitudinal and cross-sectional measures of school VA turn out to be very correlated; by contrast, the correlation between the two measures is much lower when student attrition is high. Second, notwithstanding the problem of potential non-random student attrition, we show that longitudinal models, controlling for past test scores, lead to school VAs estimates that are less sensitive to sample selection. This holds true in both high and low attrition settings, and points to the importance of testing the same cohort of students over time to build robust measures of school VAs.

Another finding in our analysis is that the persistence in test scores (the first order autoregressive coefficient) estimated in longitudinal models of school VA is higher in high attrition educational settings. We propose a rationalization of this evidence based on a simple conceptual framework where individual competences depends on past performance, own ability and abilities of the peer group, along with the variance of these abilities in the peer group. Intuitively, more selective systems, *i.e.* systems in which there is a high number of drop outs or school switchers, may be those in which individuals were initially mismatched with respect to the schools they enrolled in. This will be reflected in a higher heterogeneity in the school intake, both at a given point in time and overtime, especially in the initial grades of a school cycle (e.g., in the case of the PISA tests analysed in this paper students were sampled in the second and third year of upper secondary education). If this is the case, the school fixed effects are very poor proxies of (*i.e.*, less correlated with) students' characteristics, including past performance, inducing a higher persistence in test scores (*i.e.*, a higher first autoregressive coefficient). Moreover, a higher school mismatch also entails a peer group which changes overtime both in average ability and in its variance. As peer group's effects enter the estimate of the school VA (school fixed effects) this also implies a higher variability in cross-sectional measures of school VAs overtime in more selective schooling environments.

Thus our main policy recommendation is that VAs measures should be taken with caution in school settings where high selectivity takes place. Although this may be quite obvious in the case of curricular competences, we show that this is also the case when a measure of knowledge which should be less sensitive to school inputs, like that provided by PISA, is used. While in compulsory education in comprehensive schools one may accept the underlying assumption of students being randomly allocated to schools and teachers (despite residential segregation may work against such an assumption), in non-compulsory schooling in tracked secondary school systems, school switching and dropping-out represent an often unsurmountable obstacle for using student tests to evaluate schools performance. Repeated testing of students may represent a partial way out of the problem, as long as student attrition is not excessive.

The main limitation of our study is the reliance on single case studies, without the possibility to check whether our suggested interpretation holds for a larger variety of situations (including other geographical contexts, and/or other school grades). In the recent years the National Agency for the Evaluation of the School System (INVALSI) has developed a procedure to link student test scores over different grades (grade 2, 5, 8 and 10), which in principle makes our exercise replicable on a wider scale. However, looking at aggregate school leaving rates, school selectivity takes place between grade 8 and grade 11, where only one test measure is available. Thus ad-hoc experiments remain necessary if one aims to deepen the issue of secondary school assessment.

TABLE 1

DESCRIPTIVE STATISTICS, BY MACRO-REGIONS AND TYPE OF SCHOOL ATTENDED - ITALY PISA 2009

Literacy (mean, standard deviation and number of observations)

macroregion	high school	technical school	state vocational	regional vocational	Total
Trento	563.69	510.60	472.73	414.57	507.5
	62.69	61.77	70.35	71.98	86.49
	575	425	136	311	1,447
Valle d'Aosta	560.37	513.58	464.01	424.71	517.11
	65.33	61.37	71.78	55.63	81.34
	432	121	283	35	871
other North	562.38	508.98	459.41	414.16	507.27
Eastern regions	63.38	62.61	78.08	75.55	85.97
	2,019	1,558	947	737	5,261
other North	562.28	503.58	435.11	401.96	512.36
Western regions	64.35	67.95	82.89	73.58	88.19
	2,031	1,343	824	235	4,433
Central and	532.80	457.98	399.31	365.60	480.19
Southern regions	66.70	72.69	75.35	67.55	88.85
	8,819	5,934	3,796	219	18,768
Italy	543.56	476.08	418.47	405.70	491.78
	67.19	73.77	81.02	74.91	89.16
	13,876	9,381	5,986	1,537	30,780

Numeracy (mean, standard deviation and number of observations)

macroregion	high school	technical school	state vocational	regional vocational	Total
Trento	549.36	534.21	475.66	443.81	515.3
	71.03	60.46	72.29	63.30	78.96
	575	425	136	311	1,447
Valle d'Aosta	533.45	531.49	462.04	409.9	505.01
	77.59	66.75	68.47	49.66	81.71
	432	121	283	35	871
other North Eastern	552.99	529.96	464.09	445.45	515.10
regions	68.99	64.83	70.53	70.02	80.10
	2,019	1,558	947	737	5,261
other North	546.68	511.93	438.01	412.36	508.83
Western regions	71.07	65.34	74.46	72.00	83.63
	2,031	1,343	824	235	4,433
Central and	511.04	473.18	407.06	382.87	476.54
Southern regions	75.04	74.77	71.35	63.56	84.44
	8,819	5,934	3,796	219	18,768
Italy	524.65	491.67	424.50	430.33	490.41
	75.75	75.42	75.55	71.43	85.08
	13,876	9,381	5,986	1,537	30,780

Source: Our computations on OECD-PISA 2009 data, Italian sample.

TABLE 2

VARIANCE EXPLAINED BY SCHOOL TYPES – ITALY PISA 2009

	reading		numeracy	
	variance explained by school track fixed effect (R^2)	variance explained by school fixed effect (R^2)	variance explained by school track fixed effect (R^2)	variance explained by school fixed effect (R^2)
Trento	0.43	0.55	0.29	0.49
Valle d'Aosta	0.33	0.50	0.22	0.48
Rest of North East	0.38	0.54	0.27	0.50
Rest of North West	0.37	0.57	0.30	0.51
Centre and South	0.37	0.57	0.23	0.52
Italy	0.34	0.58	0.21	0.54

Source: Our computations on OECD-PISA 2009 data, Italian sample.

TABLE 3

FAMILY BACKGROUND BY SCHOOL TYPE: MEANS OF HIGHEST OCCUPATIONAL PRESTIGE, HIGHEST YEARS OF PARENTAL EDUCATION AND ESCS (PISA index of economic, social and cultural status) – Italy PISA 2009

macro region	high school	technical school	state vocational	regional vocational	Total
Trento	51.10	43.83	44.30	37.12	45.39
	13.99	13.01	13.64	12.28	13.31
	0.20	-0.22	-0.12	-0.61	-0.13
Valle d'Aosta	52.95	45.02	41.94	40.66	47.81
	13.86	12.7	12.23	11.2	13.07
	0.21	-0.23	-0.46	-0.73	-0.11
Rest of North East	53.92	45.25	43.33	38.42	47.33
	14.14	12.71	12.51	12.23	13.16
	0.33	-0.22	-0.36	-0.59	-0.08
Rest of North West	55.96	45.42	41.96	36.42	49.18
	14.39	12.88	12.42	11.58	13.42
	0.43	-0.18	-0.47	-0.79	0.01
Centre and South	52.55	42.66	38.65	35.77	46.48
	13.86	12.32	11.72	11.29	12.92
	0.28	-0.34	-0.65	-0.87	-0.12
Italy	53.20	43.58	40.15	37.53	47.00
	13.98	12.5	12.01	11.98	13.05
	0.30	-0.29	-0.56	-0.67	-0.09

Source: Our computations on OECD-PISA 2009 data, Italian sample.

Table 4

SCHEME FOR BOOKLET ROTATION IN THE PISA 2010 RE-TEST

Bookid	**Pisa 2009**				**Bookid**	**Re-test 2010**			
1	M1	R1	R3	**M3**	7	R6	**M3**	S3	R4
2	R1	**S1**	R4	R7	3	**S1**	R3	M2	S3
3	S1	R3	**M2**	S3	5	R4	**M2**	R5	M1
4	R3	R4	**S2**	R2	9	M2	**S2**	R6	R1
5	R4	**M2**	R5	M1	11	M3	R7	R2	**M2**
6	**R5**	R6	R7	R3	13	S3	R2	R1	**R5**
7	R6	**M3**	S3	R4	1	M1	R1	R3	**M3**
8	R2	M1	**S1**	R6	2	R1	**S1**	R4	R7
9	M2	**S2**	R6	R1	4	R3	R4	**S2**	R2
10	S2	R5	M3	**S1**	8	R2	M1	**S1**	R6
11	**M3**	R7	R2	M2	10	S2	R5	**M3**	S1
12	**R7**	S3	M1	S2	6	R5	R6	**R7**	R3
13	**S3**	R2	R1	R5	12	R7	**S3**	M1	S2

Notes: Grey cells correspond to the common sections in both 2009 and 2010 tests. Each booklet section is identified by a letter indicating the typology (M for maths, S for sciences and R for reading) and a numeric value.

Table 5

PARTICIPATION TO RE-TEST IN 2010 – VALLE D'AOSTA

→ school attended in 2010 ↓ school attended in 2009	high schools	technical school	state vocational school	regional vocational school	out of schooling/ absent/not matched	Total	% drop-out	% drop-out +mobility to different schools
high schools	352	2	24	3	51	432	0.118	0.185
technical school	0	100	5	0	16	121	0.132	0.174
state vocational school	1	0	222	3	57	283	0.201	0.216
regional vocational school	0	0	0	22	13	35	0.371	0.371
lower secondary	0	0	2	0	6	8	0.750	1.000
Total	**353**	**102**	**253**	**28**	**143**	**879**	**0.163**	**0.208**

Source: Our computations on OECD-PISA 2010 Valle d'Aosta's retest.

TABLE 6

SAMPLE MEANS OF READING TEST SCORES – STUDENTS REMAINING IN THE SAME SCHOOLS – VALLE D'AOSTA

	test in Italian		test in French	
	2009 test	2010 test	2009 test	2010 test
high school (*Licei*)	568.14	576.69	569.44	525.78
technical schools (*Istituti tecnici*)	519.74	526.29	525.04	450.81
state vocational schools (*Istituti professionali statali*)	474.70	479.73	474.97	416.74
regional vocational schools (*Centri formazione professionale regionali*)	429.15	367.77	430.16	293.20
Total	**527.34**	**532.35**	**529.20**	**473.58**

Source: Our computations on OECD-PISA 2010 Valle d'Aosta's retest data.

TABLE 7

ALTERNATIVE STRATEGIES TO IDENTIFY SCHOOL FIXED EFFECTS (intercepts only) – TRENTO

VARIABLES	(1) TN 2009 panel component no SFE	(2) TN 2009 panel component SFE	(3) TN 2010 panel component no SFE	(4) TN 2010 panel component SFE	(5) TN 2009-10 panel component SFE	(6) TN 2009-10 panel component SFE	(7) TN 0910 panel component SFE with PS
test score in 2009					0.257***	0.295***	0.304***
(reading)					[0.054]	[0.055]	[0.070]
female	4.921	-4.277	5.885	-3.647	-2.547		-1.277
	[7.401]	[7.570]	[7.953]	[7.090]	[5.944]		[5.935]
age of student	22.093**	14.365**	20.118*	4.784	1.089		-10.85
	[9.676]	[6.524]	[11.155]	[9.230]	[8.452]		[14.580]
grade compared to modal grade	52.383**	40.467***	35.144**	15.597*	5.186		18.718
in country	[19.996]	[11.325]	[14.690]	[9.076]	[8.883]		[12.913]
attended kindergarten	30.921*	20.279	23.722*	14.406	9.189		15.087
	[16.674]	[15.078]	[13.080]	[13.957]	[13.307]		[13.874]
single parent	8.244	12.706	-1.239	3.4	0.131		-15.23
	[9.751]	[9.111]	[11.480]	[11.915]	[11.572]		[15.678]
how many books	18.405***	8.397***	15.118***	3.973*	1.812		4.341
at home	[2.944]	[1.885]	[2.403]	[2.277]	[2.231]		[3.720]
highest parental	-0.99	-2.183*	-0.485	-2.039*	-1.477		-3.046
education in years	[1.157]	[1.188]	[1.072]	[1.040]	[0.969]		[1.896]
highest parental	1.086***	0.334*	0.858*	-0.108	-0.194		-0.21
occupational status	[0.257]	[0.195]	[0.451]	[0.381]	[0.366]		[0.364]
wealth	-5.612	-3.898	-7.198	-5.267	-4.265		-6.101
	[4.668]	[5.180]	[6.123]	[4.694]	[5.152]		[4.898]
second-generation	-10.144	1.262	-9.594	6.928	6.604		-8.415
immigrants	[14.291]	[14.232]	[24.709]	[29.179]	[28.190]		[28.485]

continued

continued Table 7

VARIABLES	(1) TN 2009 panel component no SFE	(2) TN 2009 panel component SFE	(3) TN 2010 panel component no SFE	(4) TN 2010 panel component SFE	(5) TN 2009-10 panel component SFE	(6) TN 2009-10 panel component SFE	(7) TN 0910 panel component SFE with PS
first-generation immigrants	-41.721*	-41.763***	-38.894*	-39.041***	-28.296**		-28.605**
	[22.370]	[14.770]	[19.946]	[11.117]	[10.725]		[10.697]
propensity score (PS)							-159.94
							[146.637]
Observations	753	753	753	753	753	753	753
R^2	0.289	0.523	0.182	0.472	0.503	0.492	0.503
Number of schools	35	35	35	35	35	35	35
School FE	NO	YES	NO	YES	YES	YES	YES

CORRELATION AMONG SCHOOL FIXED EFFECTS

Model	(2)	(4)	(5)	(7)	student FE
SFE 2009 (2)	1.000				
SFE 2010 (4)	0.918	1.000			
longitudinal SFE 2009-2010 (5)		0.865	0.993	1.000	
longitudinal SFE 2009-2010 with PS (7)		0.876	0.987	0.989	1.000
for comparison:					
longitudinal SFE 2009-2010 with student FE	-0.263	0.117	0.225	0.179	1.000

RANK CORRELATION AMONG SCHOOL FIXED EFFECTS

Model	(2)	(4)	(5)	(7)	student FE
SFE 2009 (2)	1.000				
SFE 2010 (4)	0.887	1.000			
longitudinal SFE 2009-2010 (5)	0.809	0.975	1.000		
longitudinal SFE 2009-2010 with PS (7)	0.818	0.963	0.974	1.000	
for comparison:					
longitudinal SFE 2009-2010 with student FE	-0.347	0.020	0.162	0.116	1.000

Source: Our estimates on OECD-PISA 2010 Trento's retest data.
Note: Robust standard errors in brackets clustered at school level: * significant at 10%; ** significant at 5%; *** significant at 1%. Weight = student weights.

TABLE 8

ALTERNATIVE STRATEGIES TO IDENTIFY SCHOOL FIXED EFFECTS
(intercepts only) – VALLE D'AOSTA

VARIABLES	(1) AO 2009 panel component no SFE	(2) AO 2009 panel component SFE	(3) AO 2010 panel component no SFE	(4) AO 2010 panel component SFE	(5) AO 0910 panel component SFE	(6) AO 0910 panel component SFE	(7) AO 0910 panel component SFE with PS
test score in 2009 (reading)					0.688*** [0.036]	0.778*** [0.033]	0.629*** [0.100]
female	9.012 [6.367]	9.627** [4.621]	19.890*** [6.457]	22.483*** [6.106]	15.866** [5.772]		13.731* [6.886]
age of student	22.560*** [7.006]	10.341 [7.120]	20.753** [9.532]	5.979 [9.671]	-1.135 [8.023]		2.2 [8.235]
grade compared to modal grade in country	62.584*** [8.682]	36.364*** [4.931]	67.029*** [11.685]	40.027*** [7.176]	15.015** [5.775]		4.007 [17.298]
attended kindergarten	60.945** [22.830]	32.750* [17.535]	68.407*** [20.202]	33.681* [18.998]	11.18 [18.931]		-6.884 [28.633]
single parent	-8.442 [8.908]	-2.893 [5.760]	-11.151 [10.574]	-8.323 [7.564]	-6.313 [4.896]		-6.306 [4.858]
how many books at home	13.821*** [2.639]	10.260*** [2.251]	14.724*** [2.714]	10.914*** [1.880]	3.853** [1.662]		4.302** [1.950]
highest parental education in years	0.958 [1.018]	-0.493 [0.762]	3.334** [1.285]	1.697* [0.959]	2.035** [0.748]		1.489 [1.249]
highest parental occupational status	0.679** [0.292]	-0.011 [0.193]	0.49 [0.363]	-0.209 [0.258]	-0.201 [0.166]		-0.115 [0.203]
wealth	-9.531 [6.044]	-3.34 [2.438]	-6.223 [5.871]	-1.412 [2.462]	0.889 [2.617]		0.872 [2.638]
second-generation immigrants	5.573 [45.060]	-9.68 [36.019]	75.713** [31.018]	55.108** [23.113]	61.662*** [21.253]		60.760*** [21.283]
first-generation immigrants	-32.731** [15.600]	-38.683*** [13.423]	-46.851** [16.580]	-51.842*** [12.393]	-25.236** [11.833]		-24.708* [12.028]
test conducted in French			-60.519*** [5.939]	-62.690*** [4.874]	-62.483*** [3.763]	-61.604*** [3.501]	-62.352*** [3.731]
propensity score (PS)							70.012 [101.390]
Observations	663	663	663	663	663	663	663
R^2	0.345	0.62	0.38	0.625	0.741	0.721	0.741
Number of schools	22	22	22	22	22	22	22
School FE	NO	YES	NO	YES	YES	YES	YES

Source: Our estimates on OECD-PISA 2010 Valle d'Aosta's retest data.
Note: Robust standard errors in brackets clustered at school level - * significant at 10%; ** significant at 5%; *** significant at 1%. Weight = student weights.

continued Table 8

CORRELATION AMONG SCHOOL FIXED EFFECTS

Model	(2)	(4)	(5)	(7)	student FE
SFE 2009 (2)	1				
SFE 2010 (4)	-0.0192	1			
longitudinal SFE 2009-2010 (5)	0.189	0.8494	1		
longitudinal SFE 2009-2010 with PS (7)	0.1862	0.8558	0.9962	1	
for comparison:					
longitudinal SFE 2009-2010 with student FE	0.3086	0.5808	0.9171	0.9101	1

RANK CORRELATION AMONG SCHOOL FIXED EFFECTS

Model	(2)	(4)	(5)	(7)	student FE
SFE 2009 (2)	1				
SFE 2010 (4)	-0.1338	1			
longitudinal SFE 2009-2010 (5)	0.1022	0.7222	1		
longitudinal SFE 2009-2010 with PS (7)	0.0864	0.773	0.9898	1	
for comparison:					
longitudinal SFE 2009-2010 with student FE	0.3461	0.2332	0.7538	0.7154	1

Source: Our estimates on OECD-PISA 2010 Valle d'Aosta's retest data.

Table 9

MAIN DIFFERENCES BETWEEN TRENTO AND VALLE D'AOSTA PISA 2010 RE-TESTS

Empirical "facts"	Trento	Valle d'Aosta
Correlation between 2009 and 2010 cross-sectional school VAs	High	Low
Coefficient on lagged performance in the longitudinal model of school VA	Low	High

TABLE 10

ALTERNATIVE MEASURES OF FAMILY BACKGROUND AND SCHOOL TYPE GRADIENTS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Trento				Valle d'Aosta				Valle d'Aosta (only Italian)		
	TN 2009 panel component cross-sect	TN 2010 panel component cross-sect	TN 2010 panel component longitudinal	Δ coeff	AO 2009 panel component cross-sect	AO 2010 panel component cross-sect	AO 2010 panel component longitudinal	Δ coeff	AO 2010 panel component no French cross-sect	AO 2010 panel component no French longitudinal	Δ coeff
test score in 2009 (reading)			0.348 [0.049]**				0.702 [0.033]**			0.649 [0.035]**	
grade attended (compared to modal grade)	41.174 [13.587]**	24.686 [9.573]*	10.362 [7.668]	-58.0%	48.998 [5.569]**	50.21 [7.782]**	15.788 [5.721]*	-68.6%	50.895 [7.149]**	15.711 [7.052]*	-69.1%
how many books at home	11.631 [2.186]**	7.795 [2.283]**	3.749 [2.199]	-51.9%	12.505 [2.296]**	13.301 [2.042]**	4.533 [1.752]*	-65.9%	14.691 [3.022]**	5.197 [2.584]	-64.6%
highest parental education in years	-1.748 [1.158]	-1.441 [1.073]	-0.832 [0.977]	-42.3%	-0.348 [0.781]	1.739 [1.017]	1.987 [0.758]*	14.3%	2.932 [0.903]**	3.261 [0.560]**	11.2%
first-generation immigrants	-42.234 [16.584]*	-36.927 [13.280]**	-22.234 [10.901]*	-39.8%	-36.569 [14.776]*	-52.038 [13.608]**	-26.342 [12.438]*	-49.4%	-34.888 [17.775]	-29.192 [17.527]	-16.3%
academic oriented schools (*Licei*)	116.732 [15.760]**	135.415 [14.409]**	94.806 [14.382]**	-30.0%	117.687 [13.330]**	197.632 [23.276]**	114.962 [22.931]**	-41.8%	189.965 [29.399]**	107.8 [28.728]**	-43.3%
technical schools (*Istituti tecnici*)	70.691 [14.461]**	82.227 [13.103]**	57.635 [10.774]**	-29.9%	89.419 [12.610]**	157.319 [22.073]**	94.511 [23.435]**	-39.9%	166.583 [29.329]**	107.145 [28.712]**	-35.7%
state vocational schools (*Istituti professionali statali*)	32.509 [18.399]	62.738 [18.977]**	51.429 [13.832]**	-18.0%	44.975 [16.281]*	119.683 [23.687]**	88.084 [22.754]**	-26.4%	118.577 [30.813]**	85.512 [28.023]**	-27.9%
French						-61.009 [4.824]**	-61.74 [3.764]**				
Observations	753	753	753		663	663	663		328	328	
R^2	0.45	0.38	0.45		0.52	0.57	0.72		0.53	0.69	

Source: Our estimates on OECD-PISA 2010 Trento's and Valle d'Aosta's retest data.

Note: robust standard errors in brackets clustered at school level - weight=student weights - * significant at 5%; ** significant at 1%. Regressors include gender, age, kindergarten attended, single parent, parental occupation, proxy for family wealth, second generation immigrants.

FIGURE 1

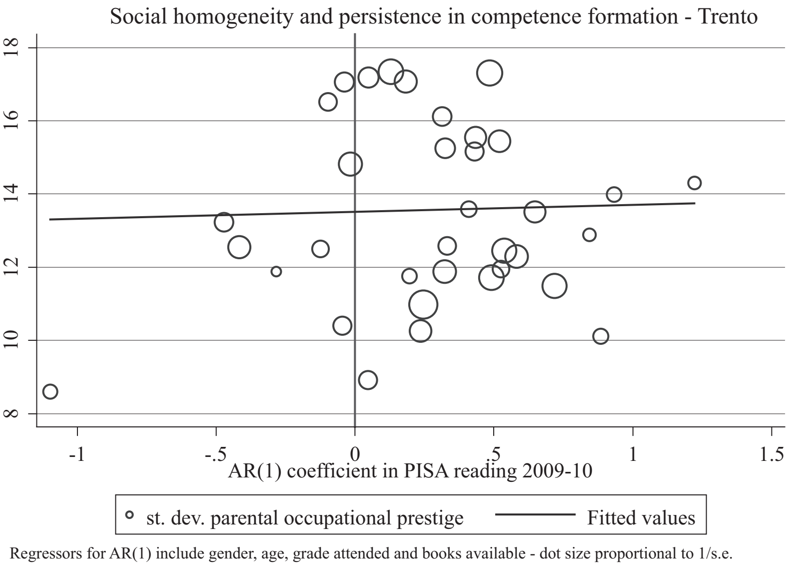
LOCATION OF THE RELEVANT PROVINCES

Valle d'Aosta (left) and Trento (right)

Source: Our elaboration.

FIGURE 2

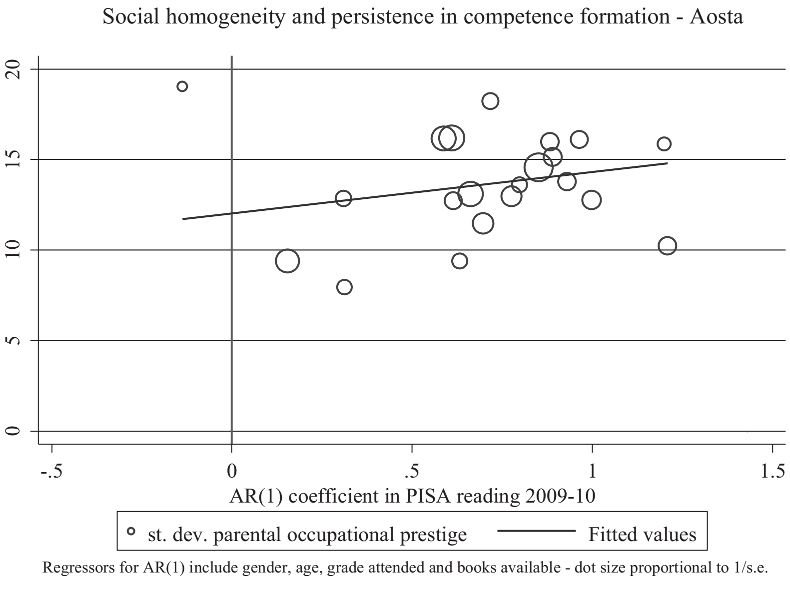
SOCIAL HOMOGENEITY AND PERSISTENCE - TRENTO



Source: Our elaboration on OECD-PISA 2010 Trento's retest data.

FIGURE 3

SOCIAL HOMOGENEITY AND PERSISTENCE - VALLE D'AOSTA



Source: Our elaboration on OECD-PISA 2010 Valle d'Aosta's retest data.

APPENDIX

A. Potential Sample Selection Biases

The two re-tests conducted in Trento and Valle d'Aosta are potentially affected by sample selection biases. The bias is likely to be more severe in Trento, as participation to the 2010 re-test took place on a voluntary basis, than in Valle d'Aosta. However, there are potential sources of bias also for the latter even if all schools participated in the re-test. In what follows we examine what factors are associated with schools' (for Trento) and students' participation in the 2010 re-tests (for both provinces).

Trento

As we have anticipated in Section 3, participation of schools to the Trento's PISA 2010 re-test took place on a voluntary basis. For this reason, the results of the re-test exercise cannot be considered as representative of the 16-year-old population of Trento. What kind of bias can be expected from the schools' self-selection into the re-test? It is plausible to think that the relatively better performing schools in PISA 2009 may have accepted to participate (*positive selection*) since they were expecting better results also in the 2010 re-test (even if they were unaware of their placement when they took the decision). In this case, the 2010 re-test may overestimate the competences of Trento's students. However, such *positive selection* is less likely to apply with respect to VA, namely the specific contribution given by schools to the improvement of student competences, as schools may have only vague notions of their VA.[26]

Among the 49 Trento's upper secondary schools sampled in PISA 2009, 14 (29%) refused to participate in the 2010 follow-up. Among them, there are 4 academic oriented schools (*licei*), 4 technical schools, 2 vocational state schools and 4 regional centres. Observing the school's distribution, there is no clear evidence of positive selection, according to which we would have expected *licei* to participate more than other school types. In Table A1 we analyse the potential distortions produced by the sample selection on the educational production function

[26] Even if the schools may ignore the VA associated to each class, for the US ROTHSTEIN J. (2009) shows evidence of non-random assignment of teachers to classes also in terms of potential competences' improvements, *i.e.* of VA.

(EPF) estimates. In column (1) we have reported a simple OLS estimation of the EPF for 2009.[27] Variables which are significantly associated with test scores are the attended grade, having attended kindergarten, books at home, highest parental occupational status, immigration status and school track, much in line with the previous literature. In column (2) we report the probit coefficients for the probability of schools' participating in the re-test. Owing to the small sample size (49 schools), we have specified a parsimonious model. Among the regressors only (school average) past kindergarten attendance, the (school average) number of books at home and the (school average) PISA 2009 test score are significantly associated to schools' participation, respectively with positive (kindergarten and books) and negative (previous year score) signs. After controlling for past performance, school track is not a significant predictor of school participation in the re-test. Past performance seems to contribute in a substantial way to the likelihood of schools' participating in the re-test, which falls by about 0.7 percent points (p.p.) for a one-point increase in the 2009 (school average) score. This means that schools which are one standard deviation (100 points) above the average performance have a 70 p.p. lower probability of participating to the re-test. Curiously enough, this indicates negative rather than positive selection. A possible interpretation of this result is that some schools which performed relatively well in 2009 preferred not to participate because they were afraid of lowering their performance in 2010 (or simply they would not spend additional instructional time, since the incentives were absent).

However, our data may be affected by a second source of selection. The schools which decided to participate might have adopted strategic behaviours by encouraging participation of abler students and discouraging that of least able students, so as to artificially inflate their measured performance.[28] In any case, voluntary absences are likely to be higher among low-performing students, which bias upward the PISA scores.[29] There is also another source of panel attrition which may potentially bias our estimates: some students may have dropped-out from educa-

[27] In all the estimates we take as measures of students' performance the average between the five plausible values provided by the OECD or INVALSI. We are aware that using simple OLS estimates without replication weights leads to biased standard errors, but it is not clear which strategy should be adopted when multiple plausible values exist both on the RHS and the LHS of the estimated equation (as in equation *(6)* for $T_{ijt\text{-}1}$ and T_{ijt}).

[28] See BRATTI M. *et* AL. (2004) for a discussion on this in the context of Higher Education.

[29] Although the direction of the bias on school VA is less clear.

tion or transferred to another school, and this is unlikely to be random with respect to their past performance. In particular, we expect least able students to be more likely to drop out or transfer to other schools, which may introduce an upward bias in the measured competences. Also in this case, like for schools' participation in the re-test, it is unlikely that the students' self-selection may have taken place based on the knowledge of the true school VA: as a consequence also this potential source of positive bias should be only minor. The percentage of absences from the re-test is 18.8%: 10.43% are ordinary absences, 0.19% refer to children whose parents denied permission, 0.10% to children with special conditions (e.g., disability), 8.04% to children who dropped out or transferred to another school. As it is clear, most absences represent ordinary student truancy, but as we said, randomness is unlikely also for these absences.

In column (3) we report the probit estimates of a student's probability to participate in the re-test exercise, conditional on her school having participated. In this model, which is estimated at the student level, we can include a wider set of controls compared to column (2). We have included all controls already considered in the estimation of the EPF in column (1). Indeed, most of these controls may have a direct effect on absenteeism, e.g., highly educated parents may value education more and push their children to reduced absenteeism, and have indirect positive effects through past (or expected future) performance. The model in column (3) also includes the day of the week in which the test was administered by the school, as student truancy may be concentred especially in certain days of the week. This variable will be also useful in our later attempt to address student self-selection in the estimation of the EPF. The results in column (3) show that there is indeed a statistically significant positive association between PISA past performance and the likelihood of participating in the re-test exercise, although the marginal effect is not very large: a one-standard deviation increase in the PISA score (100 points) raises the probability of participating by 3 p.p. Column (3) suggests that only past performance and the day of testing are strongly associated with student participation. Curiously enough, absenteeism does not appear to be related to family background after controlling for past performance. Absences turn out to be significantly more frequent on Wednesday (-9.3 p.p. in the probability of participation) and Saturday (-10.5 p.p.), *i.e.* in the middle and at the end of the week. The Wald test for the exclusion of the day of the week from the probit model returns a value of 21.5 distributed as a χ^2 (5).

Column (4) re-estimates the same model reported in column (1) restricted to the sample of students participating in the 2010 re-rest. A comparison between column (1) and (4) suggests only minor changes in the magnitude of the significant coefficients, indicating that negative selection among schools, attrition and positive selection among students in the re-test somehow compensate each other, reducing the bias in the relevant coefficients. The statistical significance of the coefficient of the dummy for vocational schools is reduced, suggesting that vocational schools may have pushed only their better students to participate in the re-test.

Column (5) reports the model including the same regressors and estimated on the sample of students who participated in both PISA tests (the panel component) but considering the score in the PISA 2010 reading test as the dependent variable. Here we note some non-negligible changes with respect to the previous column. In particular, the effect of the school grade almost halves, while that of state vocational school doubles. In general, regional vocational schools (the excluded case) lose ground with respect to all the other school types. These effects are likely to be associated with the differential retention policies and drop-out in the different schools, and are consistent with regional vocational schools being the least selective. The difference in the average ability between regional vocational schools and the other school types tend to increase at higher grades).

In column (6) we make an attempt to correct the EPF's estimates for student self-selection. In particular, we include the propensity score (PS), the probability of attending the 2010 re-test obtained from the probit model in column (3) (see Angrist 1997). The coefficient on the propensity score is positive but significant only at the 10% level, and the coefficients associated with statistically significant regressors do not change much with respect to column (5). The only exceptions are the proxy for cultural capital (number of books at home) and student grade which lose statistical significance presumably owing to their correlation with the propensity score (*i.e.* the probability of participating in the re-test).

In column (7) and (8) we follow an alternative procedure to address the selectivity issue. We estimate a Heckman sample selection model where the EPF is reported in column (7) and the selection equation in column (8).[30] The selection equation uses the same specification as in column (3). The results in column (7) are very similar to those in column (6). The estimated correlation between the error terms of the EPF and the selection equation turns out to be high in magni-

[30] The model was estimated in one step with maximum likelihood.

tude (-0.886) and very significant (the standard error is 0.046), confirming again that relatively better performing students participated in the 2010 re-test.[31]

Summing up, the analysis for Trento suggests that two countervailing forces were at play. On the one hand schools with a better performance in 2009 were less likely to participate in the 2010 re-test. On the other hand, conditional on school participation, relatively better students took part in the re-test. Thus, controlling for past student performance becomes important because it helps reduce the distortions produced by self-selection into the re-test.

Valle d'Aosta

Unlike for the re-test conducted in Trento, the 22 schools of Valle d'Aosta were obliged to administer to their entire student body the PISA test in both years. However, grade 10 (which is the modal grade for 15-year-old students) concludes compulsory education in Italy, and a fraction of students abandoned their educational career, while another fraction kept on but changed schools. As we have anticipated in Section 3, school leavers and movers jointly represent 21% of the student body, and we wonder how this may affect the estimate of schools' VA. School leavers are likely to be weaker students, thus raising the average test performance of the remaining students. If track allocation of students is not random, school switchers (mostly from academic oriented tracks to technical and vocational schools) are likely to raise the performance of the abandoned schools as well as of the receiving schools.

Following the same scheme used for the Trento re-test, Table A2 analyses the potential distortions produced by sample selection on the EPF estimates. In column (1) we have reported a simple OLS estimation of the EPF for 2009. This cross-section EPF is consistent with theoretical expectations (as well as with pre-

[31] Indeed, let us write the main performance equation as $T_{ijt} = X_{ijt}\beta + \varepsilon_{ijt}$ and the selection equation as $P_{ijt} = Z_{ijt}\delta + u_{ijt}$ where Z_{ijt} is a vector of regressors including all covariates in the vector X_{ijt} and an excluded variable used to identify the model, P_{ijt} is an indicator variable which takes value one in case of participation in the re-test exercise and zero otherwise, and ε_{ijt} and u_{ijt} two normally distributed error terms. It can be shown that $E(T_{ijt}|P_{ijt}) = X_{ijt}\beta + \rho\sigma\lambda(Z_{ijt}\delta) + \varepsilon_{ijt}$ where ρ is the correlation coefficient between ε_{ijt} and u_{ijt} and σ is the standard deviation of ε_{ijt}. $\lambda(Z_{ijt}\sigma) = \phi(Z_{ijt}\delta)/\Phi(Z_{ijt}\delta)$ is the so called Inverse Mill's (IMR) ratio, where $\phi(.)$ and $\Phi(.)$ are the standard normal density and distribution functions (see HECKMAN J.J., 1979). Thus the quantity on the denominator of the IMR is the propensity score, and $\rho < 0$ implies a positive dependence of test scores on the propensity score. Thus, results in column (7) are consistent with column (6).

vious results for Trento): test score in literacy is positively correlated with grade (repeating students have a lower score, while early beginners, *i.e.* those enrolled before the legal age, do better, as both variables are likely to capture ability), books at home (which is probably very correlated with parental years of education, which are not statistically significant) and first-generation immigrant status, which is negatively associated with reading ability in the host country. There are positive *premia* in test scores associated with attending an academic oriented school (*liceo*) compared to the excluded category of regional vocational schools (100 test points, equivalent to one standard deviation, and in line with the corresponding estimate for Trento) or a technical school (72 test points); no difference is observed between state-organised or region-organised vocational schools. These results are in line with the previous literature, suggesting that a large fraction of school differences is captured by student sorting into tracks.

In column (2) we report the probit coefficients for the probability of students' remaining in the regional schooling system, irrespective of the school attended (724 over 839 students with non-missing information), whereas column (3) restricts the sample to the students who remained in the same school (663 students[32]), thus controlling for both the probability of dropping-out and moving downward in the school tracks. Both columns indicate that there is a positive selection into taking the test the following year, based on past performance, with the marginal effect in the former (0.05 additional probability points per one hundred points in test score) being almost half of the impact measured in the latter (0.09 additional probability points per one hundred points in test score). More surprising is that the leaving/moving probability is independent of school track, suggesting that the downward mobility (from academic oriented to vocational schools) of students partially offsets the differential in dropping out (see the final two columns of Table 5). Some evidence of a role of family resources is reflected in the negative correlation with parental occupational status, which however does not predict school change at a large extent. Differently from the case of Trento, we cannot exploit additional information on test taking, like the week day of the test (since all students were tested on the same day), which would help us to iden-

[32] Let us remind the sample selection leading to the panel component in Valle d'Aosta (see Section 3): from 879 students taking the test in 2009, only 736 took the test in 2010, with 40 changing track and 12 changing schools within the same track. Thus the panel component consists of 684 students, which falls to 663 because of missing information on some demographic variables.

tify the effect of the propensity score whose identification then depends on past performance score only.[33]

Column (4) reports the estimates of the model reported in column (1), restricted to students who remained in the same school. A comparison between columns (1) and (4) indicates that there is limited selection bias in estimating the family's impact, since the correlation of performance with the number of books at home, wealth and migration status are not statistically different in the two models. Confidence intervals also overlap for dummies of school tracks, which absorb most of the parental background effect. Interestingly, the coefficient on grade attended (which captures the effect of school repeaters) declines, since part of the repeaters (probably the least able) has either dropped out or changed school.

The PISA 2009 gradients estimated in column (4) are to be compared with the PISA 2010 gradients estimated in column (5), where the sample is restricted to the students remaining in the same school.[34] The EPFs estimated in two adjacent years look rather stable. However, as in the case of Trento, we observe an increase in the magnitude of the coefficients associated with the type of secondary school attended, consistently with the idea that remaining in the school for an additional year strengthens its impact. The difference between these coefficients may be taken as gross estimates of the VA associated with the school type: thus attending an academic school in Valle d'Aosta was associated to 117 PISA points in reading in 2009 and 197 points one year later, yielding a measure of 80 points of VA. Similarly, we obtain 68 points for technical schools and 75 for state vocational schools, which are to be compared with the corresponding measures for Trento (see columns (4) and (5) in Table A1): 21 points for academic schools, 12 points for technical schools and 30 points for state vocational schools. These simple calculations highlight an issue discussed in the paper: all cross-sectional measures are conditional on a benchmark (the excluded case) and cannot be taken in their absolute value. The higher VA measures recorded in Valle d'Aosta are the likely reflection of a decline in the absolute results of regional vocational schools, especially when compared with tests administered in French. If we observe the mean test scores reported in Table 6 we notice an improvement by all school types but regional vocational schools when the test is administered in Ital-

[33] Indeed, since in Table A2 we do not adopt a longitudinal VA model (including past test score in the EPF), past test score can be used to identify the model. In longitudinal VA models, by contrast, identification will rely on functional form only.

[34] The file obtained from ACER-OECD also contains new 16-year-old entrants in 2010, which however were not made available to us due to lack of a privacy consensus agreement.

ian, and a general worsening when the test is taken in French. This is partially controlled for by a dummy variable, which suggests a penalization of almost 60 points associated with the use of French in the questionnaire, despite the strong emphasis on bilingual education in the region.

The coefficients on family background and school type change when we account for sample self-selection, either by including the estimated probability of entering the permanent component of the sample (from column (3)) in column (6) or by adopting an Heckman selection model (columns (7) and (8)). In both cases, the identification relies on imposing the exclusion of 2009 test score from the prediction of 2010 test score (*i.e.* adopting a cross-sectional specification). The magnitude of the coefficients on school types falls by approximately 20 points, indicating that part of the measured VA should be imputed to student self-sorting in the same schools. The coefficient on the PS in column (7) is high and statistically significant, suggesting that increasing the probability of participating in the 2010 re-test by 10 percent points is associated with a 67-point increase in the PISA score. Surprisingly enough, the correlation coefficient between the error terms of the EPF and the selection equation in the Heckman's selection model (column (8)) turns out to be almost identical in magnitude to the Trento's experiment (-0.885 with a standard error of 0.033).

Summarising, this appendix suggests that school VA measurement in Valle d'Aosta's schools is likely biased by changes in sample composition: if school leavers and school switchers are negatively selected, we should observe an overestimate of the whole school/teacher contribution to the improvement of students' test scores. In the paper, we do nonetheless our best to obtain robust school VA measures given the limitations of the available data.

Table A1

POTENTIAL SAMPLE DISTORTIONS – TRENTO

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	TN 2009 all available obs no SFE	TN selection equation to PISA 2010 (by school)	TN selection equation to PISA 2010 within participating school (by student)	TN 2009 panel component no SFE	TN 2010 panel component no SFE	TN 2010 panel component no SFE corrected with PS	TN 2010 participating school no SFE	Heckman selection equation (by student)
female	0.996	-0.308	0.033	-12.204	-14.105	-14.463	-13.263	0.098
	[6.674]	[1.020]	[0.110]	[6.846]*	[7.511]*	[6.030]**	[8.457]	[0.114]
age of student	7.51	-1.048	-0.294	11.455	8.628	16.781	15.3	-0.23
	[6.031]	[2.772]	[0.170]*	[8.572]	[11.117]	[10.756]	[12.652]	[0.149]
grade compared to modal grade in country	44.482	-0.907	0.294	41.174	24.686	14.008	13.108	0.101
	[9.331]***	[1.283]	[0.169]*	[13.587]***	[9.573]**	[9.321]	[10.251]	[0.177]
attended kindergarten	37.554	11.93	0.15	25.593	17.143	11.915	10.649	0.07
	[10.210]***	[3.806]***	[0.249]	[15.107]*	[14.811]	[15.626]	[18.741]	[0.248]
single parent	10.123		-0.341	9.997	0.899	13.088	12.085	-0.32
	[6.258]		[0.159]**	[9.554]	[10.791]	[11.376]	[9.864]	[0.148]**
how many books at home	11.027	1.759	0.063	11.631	7.795	5.16	5.504	0.017
	[1.528]***	[0.900]*	[0.042]	[2.186]***	[2.283]***	3.273]	[2.713]**	[0.043]
highest parental education years	-1.449	-0.076	-0.038	-1.748	-1.441	-0.13	-0.238	-0.021
	[0.793]*	[0.417]	[0.021]*	[1.158]	[1.073]	[1.450]	[1.001]	[0.020]
highest parental occupational status	0.477	0.01	0	0.423	0.128	0.125	0.161	0.002
	[0.145]***	[0.075]	[0.005]	[0.202]**	[0.356]	[0.289]	[0.299]	[0.003]
wealth	-6.071		-0.047	-4.28	-5.693	-3.846	-4.166	-0.06
	[3.391]*		[0.092]	[4.620]	[5.418]	[5.134]	[4.672]	[0.078]
second-generation immigrants	-4.694		-0.313	-4.905	-4.619	8.998	4.556	-0.392
	[12.599]		[0.357]	[16.583]	[26.031]	[25.981]	[26.545]	[0.306]
first-generation immigrants	-28.147		0.031	-42.234	-36.927	-33.658	-36.996	0.018
	[10.698]**		[0.255]	[16.584]**	[13.280]***	[13.345]**	[15.751]**	[0.265]
high school (*Licei*)	109.527	1.164	0.292	116.732	135.415	121.684	120.156	-0.114
	[11.405]***	[1.335]	[0.159]*	[15.760]***	[14.409]***	[12.175]***	[14.819]***	[0.166]
technical schools (*Istituti tecnici*)	73.454	1.554	0.092	70.691	82.227	76.96	76.417	-0.115
	[9.962]***	[1.081]	[0.121]	[14.461]***	[13.103]***	[9.824]***	[12.343]***	[0.125]
state vocational schools (*Istituti professionali statali*)	38.915	-0.083	-0.01	32.509	62.738	61.345	58.832	-0.199
	[14.234]***	[0.860]	[0.106]	[18.399]*	[18.977]***	[11.325]***	[18.063]***	[0.109]*
test score 2009 (reading)		-0.026	0.001					0.003
		[0.012]**	[0.001]**					[0.001]***
day of 2010 testing = Monday			0.079					-0.08
			[0.108]					[0.151]
day of 2010 testing = Tuesday			0.136					-0.13
			[0.100]					[0.095]

continued

continued TABLE A1

	(1) TN 2009 all available obs no SFE	(2) TN selection equation to PISA 2010 (by school)	(3) TN selection equation to PISA 2010 within participating school (by student)	(4) TN 2009 panel component no SFE	(5) TN 2010 panel component no SFE	(6) TN 2010 panel component no SFE corrected with PS	(7) TN 2010 participating school no SFE	(8) Heckman selection equation (by student)
day of 2010 testing			-0.359					-0.23
= Wednesday			[0.151]**					[0.169]
day of 2010 testing			-0.089					-0.258
= Friday			[0.111]					[0.109]**
day of 2010 testing			-0.407					-0.303
= Saturday			[0.106]***					[0.119]**
probability of attending test in 2010						135.285		
(from col.3)— PS						[69.464]*		
ρ								-0.886***
								[0.046]
Observations	1,359	49	941	753	753	753	941	941
R^2/Pseudo R^2	0.45	0.21	0.06	0.45	0.38	0.38	-	-

Source: Our estimates on OECD-PISA 2010 Trento's retest data.

Note: Robust standard errors in brackets clustered at school level, weighted by student weights: * significant at 10%; ** significant at 5%; *** significant at 1%. Column 2: *probit* model for participating to PISA 2010 – column 3: probit model for participating to PISA 2010, conditional on remaining in the same school – columns 1, 4, 5, 6: OLS – columns 7-8: MLE Heckman selection model (ρ is the correlation coefficient between the errors in the EPF and the selection equations). Column (6) reports bootstrapped standard errors (1,000 replications), as the propensity score (PS) is a generated regressor.

TABLE A2

POTENTIAL SAMPLE DISTORTIONS – VALLE D'AOSTA

	(1) AO 2009 all available obs no SFE	(2) AO selection equation to PISA 2010 no SFE	(3) AO selection equation to panel component of PISA 2010 no SFE	(4) AO 2009 panel component no SFE	(5) TN 2010 panel component no SFE	(6) AO 2010 panel component no SFE corrected with PS	(7) AO 2010 participating school no SFE	(8) Heckman selection equation
female	7.564	0.083	0.135	5.11	15.494	-8.616	11.486	0.016
	[5.683]	[0.108]	[0.075]*	[5.463]	[5.953]**	[4.489]*	[5.556]**	[0.083]
age of student	5.877	-0.492	-0.206	13.222	7.214	32.789	9.596	-0.227
	[5.021]	[0.229]**	[0.223]	[7.275]*	[11.093]	[7.030]***	[13.624]	[0.202]
grade compared to modal grade in country	55.232	0.305	0.553	48.998	50.21	-90.018	29.297	0.2
	[5.535]***	[0.200]	[0.088]***	[5.569]***	[7.782]***	[11.147]***	[6.826]***	[0.104]*
attended kindergarten	14.877	0.605	0.807	37.03	40.114	-165.589	3.443	0.724
	[12.775]	[0.345]*	[0.348]**	[21.181]*	[18.912]**	[22.193]***	[28.116]	[0.349]**
single parent	-4.879	-0.269	-0.256	-9.612	-14.163	38.140	-4.698	-0.222
	[6.035]	[0.212]	[0.185]	[6.701]	[8.258]	[6.952]***	[8.005]	[0.182]
how many books at home	12.657	0.01	-0.022	12.505	13.301	9.696	12.592	-0.092
	[2.250]***	[0.033]	[0.035]	[2.296]***	[2.042]***	[1.708]***	[2.129]***	[0.037]**
highest parental education in years	-0.76	0.03	0.027	-0.348	1.739	-2.829	0.792	0.023
	[0.703]	[0.021]	[0.022]	[0.781]	[1.017]	[0.834]***	[1.019]	[0.018]
highest parental occupational status	0.147	-0.01	-0.004	0.228	-0.036	0.532	0.023	-0.005
	[0.157]	[0.003]***	[0.003]	[0.184]	[0.246]	[0.167]***	[0.219]	[0.003]*
wealth	-10.515	-0.068	0.003	-10.256	-7.333	-1.916	-6.912	0.052
	[3.749]**	[0.132]	[0.077]	[4.200]**	[3.562]*	[3.085]	[4.124]*	[0.080]
second-generation immigrants	-24.055	-0.164	0.276	-23.561	42.457	1.367	34.524	0.377
	[43.177]	[0.524]	[0.527]	[50.152]	[32.084]	[27.112]	[31.948]	[0.538]
first-generation immigrants	-41.803	0.065	0.461	-36.569	-52.038	-110.779	-61.488	0.727
	[10.609]***	[0.229]	[0.238]*	[14.776]**	[13.608]***	[12.824]***	[15.677]***	[0.220]***
high school (*Licei*)	100.557	0.56	-0.159	117.687	197.632	139.945	175.148	-1.001
	[15.475]***	[0.198]***	[0.210]	[13.330]***	[23.276]***	[18.374]***	[28.712]***	[0.391]**
technical schools (*Istituti tecnici*)	72.677	0.65	0.211	89.419	157.319	57.282	127.879	-0.54
	[12.431]***	[0.220]***	[0.292]	[12.610]***	[22.073]***	[19.235]***	[28.444]***	[0.446]
state vocational schools (*Istituti professionali statali*)	25.837	0.696	0.314	44.975	119.683	31.605	92.144	-0.139
	[15.225]	[0.174]***	[0.186]*	[16.281]**	[23.687]***	[18.553]*	[29.357]***	[0.393]
test score in 2009 (reading)		0.003	0.004					0.009
		[0.001]**	[0.001]***					[0.001]***
test conducted in French					-61.009	-60.444	-59.434	
					[4.824]***	[4.124]***	[4.182]***	
probability of attending test in 2010 (from col.3) — PS						673.1		
						[48]***		

continued

continued TABLE A2

	(1) AO 2009 all available obs no SFE	(2) AO selection equation to PISA 2010 no SFE	(3) AO selection equation to panel component of PISA 2010 no SFE	(4) AO 2009 panel component no SFE	(5) TN 2010 panel component no SFE	(6) AO 2010 panel component no SFE corrected with PS	(7) AO 2010 participating school no SFE	(8) Heckman selection equation
ρ								-0.885*** [0.033]
Observations	839	839	839	663	663	663	839	839
R^2/Pseudo R^2	0.53	0.10	0.11	0.52	0.57	0.70		

Source: Our estimates on OECD-PISA 2010 Valle d'Aosta's retest data.
Note: Robust standard errors in brackets clustered at school level, weighted by student weights: * significant at 10%; ** significant at 5%; *** significant at 1%. Column 2: *probit* model for participating to PISA 2010 – column 3: probit model for participating to PISA 2010, conditional on remaining in the same school – columns 1, 4, 5, 6: OLS – columns 7-8: MLE Heckman selection model (ρ is the correlation coefficient between the errors in the EPF and the selection equations). Column (6) reports bootstrapped standard errors (1,000 replications), as the propensity score (PS) is a generated regressor.

BIBLIOGRAPHY

Angrist J.D. (1997), «Conditional Independence in Sample Selection Models», *Economics Letters,* vol. 54, no. 2, pp. 103-112 .

American Statistical Association (2014), «Statement on Using Value-Added Models for Educational Assessment», April 8, *http://www.amstat.org/policy/pdfs/asa_vam_ statement.pdf*

Benabou R. (1996*a*), «Heterogeneity, Stratification, and Growth: Macroeconomic Implications of Community Structure and School Finance», *American Economic Review*, vol. 86, no. 3, pp. 584-609.

-.- (1996*b*), «Equity and Efficiency in Human Capital Investment: The Local Connection», *Review of Economic Studies,* vol. 62, no. 2, pp. 237-264.

Bratti M. - Checchi D. - Filippin A. (2007), «Geographical Differences in Italian Students' Mathematical Competences: Evidence from PISA», *Giornale degli Economisti e Annali di Economia*, vol. 66, no. 3, pp. 299-335.

Bratti M. - McKnight A. - Naylor R. - Smith J. (2004), «Higher Education Outcomes, Graduate Employment and University Performance Indicators», *Journal of the Royal Statistical Society Series A,* vol. 167, no. 3, pp. 475-496.

Di Chiacchio C. - Giangiacomo P. - Palmerio L. (2010), *Progetto sull'analisi del valore aggiunto - Trentino - Rapporto INVALSI sullo scoring degli Studi 2009-2010, Nota Metodologica*, mimeo, Frascati.

Heckman J.J. (1979), «Sample Selection Bias as a Specification Error», *Econometrica,* vol. 47, no. 1, pp. 153-161.

Invalsi (2016), *Rilevazioni nazionali degli apprendimenti 2015-16*, Roma, Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione. 215-16. Rome

Lazear E.P. (2001), «Educational Production», *The Quarterly Journal of Economics*, 116(3), pp. 777-803.

Oecd (2010), *PISA 2009 Results: What Students Know and Can Do - Student Performance in Reading, Mathematics and Science*, vol. I, mimeo.

Pons X. (2011), «What Do We Really Learn from PISA? The Sociology of Its Reception in Three European Countries (2001-2008)», *European Journal of Education,* vol. 46, no. 4, pp. 540-548.

Puhani P.A. (2000), «The Heckman Correction for Sample Selection and Its Critique», *Journal of Economic Surveys*, vol. 14, no. 1, pp. 53-68.

Rothstein J. (2009), «Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables», *Education Finance and Policy*, vol. 4, no. 4, pp. 537-571.

-.- (2010), «Teacher Quality in Educational Production: Tracking, Decay and Student Achievement», *Quarterly Journal of Economics*, vol. 125, no. 1, pp. 175-214.

Todd P. - Wolpin K.I. (2003), «On the Specification and Estimation of the Production Function for Cognitive Achievement», *Economic Journal*, vol. 113, no. 485, pp. F3-F33.