

EVALUATING THE STABILITY OF SCHOOL PERFORMANCE ESTIMATES FOR SCHOOL CHOICE: EVIDENCE FOR ITALIAN PRIMARY SCHOOLS

Tommaso Agasisti, Politecnico di Milano School of Management

Veronica Minaya, Politecnico di Milano School of Management

JEL Classification: I21

Keywords: accountability, school choice, school value-added

Evaluating the Stability of School Performance Estimates for School Choice: Evidence for Italian Primary Schools

Veronica Minaya

Politecnico di Milano School of Management

veronicamilagros.minaya@polimi.it

Tommaso Agasisti

Politecnico di Milano School of Management

tommaso.agasisti@polimi.it

December 2017

Abstract. School performance estimates have been used worldwide for both high-and low-stakes accountability purposes. It is expected that by evaluating school performance and making these results public, parents will use them to choose schools and schools will be motivated to increase performance. An institutional debate of this kind is likely to start in the near future in Italy, given the growing availability of indicators obtained through standardized test scores in reading and mathematics. Using administrative data provided by INVALSI (National Evaluation Committee for Education), this paper explores the stability of performance estimates for Italian primary schools. We first construct school performance metrics using INVALSI standardized tests and quarterly teacher assessments, by taking advantage of a rich array of individual level variables (including prior achievement) that allow us to estimate a school-effect in a ‘value added’ perspective. We then explore how sensitive school ratings are to the choice of performance metric and the use of different models to account for compositional differences due to students’ socioeconomic background. We also assess both cross-sectional differences in performance across schools and the persistence of these differences across cohorts. We find that school performance estimates are very robust whatever the models employed to control for compositional differences, but they are inconsistent across metrics and cohorts. We conclude that before using this kind of school-effects’ estimates for school choice purposes, more data and research is needed to understand the factors that drives the impact of a specific school on students’ results.

Keywords: accountability, school choice, school value-added

JEL Codes. I21

Acknowledgments. The research has been funded by Politecnico di Milano under the FARB Project “Public Management Research: Health and Education Systems Assessment”. We are grateful to the INVALSI’s statistical office for having provided the dataset. Previous versions of this work have been presented at a Seminar at Politecnico di Milano School of Management (Milan, Italy), at the Scientific Workshop of the H2020 Project EdEN (Budapest, Hungary) in October 2017 and at the 5th International School Choice and Reform Conference in Fort Lauderdale (FL, USA) in January 2018. We thank all the colleagues who participated to these seminars for their useful comments. All eventual errors are our solely responsibility.

1. Introduction

School performance estimates have been used worldwide for both high-and low-stakes accountability purposes. It is expected that by evaluating school quality and making these results publicly available, parents will make more informed choices about schools. Also, performance indicators can be used by school principals and teachers to reform their leadership and teaching practices for improving educational outcomes. In many countries is the norm to use school performance estimates without other specific stakes, as the so-called report-card accountability. The publication of school performance estimates is well established in some countries like in the US and the UK that have a long tradition in the production of league tables on a yearly basis. These countries also have in place high-stake accountability systems and use school performance estimates to generate stronger financial incentives and sanctions for failing to meet achievement benchmarks.

The design of accountability systems with low-versus high-stakes testing have been debated in the literature (Jacob, 2002; Figlio & Rouse, 2005; Figlio & Lucas, 2004; Figlio & Kenny, 2009; Holbein & Ladd, 2017). While some argue that accountability systems with low stakes will not generate incentives relevant enough to improve educational practice, others consider that the typical metrics used in high-stakes testing do not capture the main objectives that schools are trying to accomplish (Deming & Figlio, 2016). Standardized tests do not measure performance in other academic areas and more abstract and long-term goals such as science or critical thinking, and school completion or college enrollment. The policy experience in other countries reveal that even the choice of alternative indicators for measuring the schools' value-added is subjected to discretionary choices, that can reflect specific priorities or policy objectives. For example, Leckie & Goldstein (2017) describe the evolution of the various measures used by the British government between 1992 and 2016 to compile the schools' performance' league tables. Overall, the indicator that calculates 'school progress' has been changed four times: from valued-added (2002) to contextual value-added (2006), to expected progress (2011) to a different measure of progress (2016).

An institutional debate about the use of school performance metrics is likely to start in the near future in Italy, given the growing availability of information obtained through standardized tests for reading and mathematics. The National Evaluation Committee for Education (INVALSI) has recently started to build up a national system of school evaluation by administering annual tests in various grades (2,5,8,10). Since 2016, INVALSI has been estimating 'school effects' using a value-added approach and adjusting school performance estimates for students' socioeconomic background. These results are shared with school principals and other key stakeholders. Also, aggregate statistics about value-added have been published by the most important media and newspapers in the country. As INVALSI continues estimating school performance, policymakers may wonder how this information can be further used to improve educational outcomes if these results were accessible to parents to make school choices.

Using administrative data provided by INVALSI (National Evaluation Committee for Education), this paper explores the stability of school performance estimates for school choice for Italian primary schools. Our goal is not to isolate causal 'school effects', but to explore how

sensitive school ratings may be to the inclusion of adjustments for compositional differences, the choice of metric and over time across cohorts. The main idea behind our approach is to test whether the ratings generated for the i -th school in the t -th time period, using a certain set of data, can be considered a good predictor of this school's performance in a different year and/or with different metrics. This conceptual idea is of extremely importance when turning the attention to the potential use within school policies. Informed parents seek those schools that are able to maximize the benefits (related to objectives) for a given cost. If value added estimates reflect these benefits, parents assume that there should be some school-specific features that have an effect on the observed performance. But, if these estimations fluctuate over time, the same informed parents could believe that there is not actually a "school-effect", and that the performance obtained in a given year is due to variables that are not under the control of schools' discretions and they cannot be replicated in future.

This paper comes at an important time in Italy as policymakers debate how to design their own accountability system and to measure and communicate school performance to the public. As some researchers suggest, once you start measuring and tracking performance, it is very hard to stop doing it (Deming & Figlio, 2016). Therefore, exploring the stability of school performance estimates might help to better understand what can be reasonably expected to accomplish with the available information. Our work relates to a growing literature measuring the stability of school performance estimates across different models, performance metrics, and cohorts (Papay, 2011; Kane & Staiger, 2002; Dumay et al., 2013; Kortelainen et al., 2016; Perry; 2016). As noted by Perry (2016), a relatively small number of studies have addressed the issue of school performance stability over time and across different cohorts in the same school. We are the first to use Census data of three cohorts of Italian students enrolled in 5th grade to assess stability of school performance estimates.

First, we construct school performance metrics using INVALSI data and taking advantage of a rich array of individual level variables (including prior achievement), that allow us to estimate school effects using a 'value added' approach. We then explore how sensitive school ratings are to the use of different models to account for compositional differences and the choice of performance metric. We also assess both cross-sectional differences in performance across schools and the persistence of these differences across cohorts. Drawing on a multi-cohort longitudinal database (three cohorts, from 2013 to 2015), we especially address the issue of school performance stability over time.

The empirical analysis yields to six main findings. First, we find that even though adjusted and unadjusted estimates are very highly positively correlated, controlling for compositional differences matters for school choice. Second, differences in school performance estimates are not substantially large across schools, particularly among schools that are ranked in the middle of the distribution. Third, school performance estimates are unstable and correlations across cohorts decline over time. This result also means that it is difficult to predict the future performance of schools based on their past and current performance. Fourth, adjusted school performance correlations across cohorts have lower stability than unadjusted correlations. Fifth, school performance estimates using teachers' scores are far from perfectly correlated with those using standardized test scores. Finally, school ratings using teachers' scores have slightly higher levels

of stability compared to those based on standardized test scores. These findings have some important policy and managerial implications.

We conclude that published school performance estimates for school choice are likely to be misleading, particularly for those who are unfamiliar with the level of stability in these estimates, and that the development of a more complete database at school level – accompanied by more research – is a prerequisite before using this data for school choice and accountability purposes. For example, the current administrative datasets lack even basic information about school resources, or principals and teachers' characteristics; all of these dimensions have probably a role in affecting the level and the dynamics of the estimated school effects.

The remainder of the paper is structured as follows. Section 2 presents a policy background on school accountability, and Section 3 describes our data and sample, together with the empirical methodology. Section 4 presents the results. Section 5 concludes with broader implications.

2. Policy and institutional background

The push for school accountability, through the Italian school reform introduced in 2001, arose from well-known concerns about the performance of the Italian education system, when comparing Italian students' achievement with peers in other OECD countries. Italy performs below the average for OECD countries in mathematics, science and reading, even though is one of the countries with the largest improvement in mathematics and science performance between 2006 and 2009 (OECD, 2012). Recent evidence suggests that such improvements are slowing down, therefore in 2012 and 2015 (OECD, 2016).

As part of the efforts to make schools more responsive to students and parents' educational needs, and to enable school principals and the School Board to be held accountable for the efficient and effective use of resources (Ministry of Education, 2001), the Italian government has boosted the activity of the INVALSI, which yearly evaluates the efficiency and effectiveness of the Italian Educational System by assessing students' achievement in mathematics and reading at different stages: at the end of the second and fifth grade of the primary school (about at age 7 and 10, respectively), at the eighth grade, i.e. the end of lower secondary school (age 13) and at the end of tenth grade of upper secondary school (age 15). INVALSI tests are low stakes for all grades in primary school, as they are neither used for economic incentives to the schools or teachers, nor are considered for assigning students' grades or for deciding passing the levels. Thus, the Ministry of Education has decided to do not reveal results from INVALSI tests to the public, so at present the scores are only available to schools' internal decision-makers (principal and school board, essentially). Therefore, one current institutional debate is about the opportunity for a different – more public – use of the information generated through school-averages of the INVALSI test scores. Until now (2017), the results are only disclosed in an aggregated form (by the 21 Italian Regions) and discussed for national-level political dialogue. The so-called "National Report" with aggregated results is presented yearly to the media, and then made available in the INVALSI's website.

In some countries, within the context of accountability and school choice, school performance estimates are instead used to rate schools, in the form of ‘league tables’, to encourage informed parents’ school choices and stimulate school improvement. Evidence suggest that parents and schools actually use this information (Hasting & Weinstein, 2008; Burgess et al., 2015) and respond to accountability ratings. Parents are willing to pay more for houses located nearby high-performing schools (Black, 1999; Black & Machin, 2011) and to contribute more or less to schools through parent-teacher organizations based on accountability measures (Figlio & Lucas, 2004). Following-up analyses of their own results, schools reorganize their practice by changing the learning environment or increasing resources available to teachers (Figlio, 2013). However, researchers have also found some evidence on unintended negative school responses due to accountability pressures, such as changing the rules around student testing exemptions, teaching to the test, focusing excessively on short-term outcomes, narrowing their curriculum or instruction practices at the expense of non-tested groups or subjects, and outright teacher cheating. Deming and Figlio (2016) draw some important lessons from accountability in the US, which is the country where the related debate is the oldest.

Given the use and strategic responses to accountability pressures, it is crucial to avoid noisy or unstable metrics that would fail to predict or represent future school performance and elicit any performance response. This is also the explicit rationale advocated by the Italian Ministry of Education to maintain confidentiality about the test scores at school level, at least until the performance measures are not proven to be stable and reliable. In this vein, the academic literature concerning the bias and stability of school performance estimates have been mostly focused on assessing how stable ‘school effects’ are over different models, testing instruments and time frames.

Yet, there is still wide discussion on how calculating ‘school effects’ (Dumay et al., 2012). The methodological discussion about the various problems behind the definition of appropriate modelling choices is summarized in the recent work by Everson (2017). The main challenge researchers face in estimating ‘school effects’ is that students are not assigned to schools randomly. The most common problem is known as vertical selection, through which differences in students’ socioeconomic backgrounds may cause high or low school performance estimates. For instance, vertical selection would overestimate ‘school effects’ if parents positively choose schools for their children and to underestimate ‘school effects’ for students who are negatively assigned to schools (i.e. they are assigned to low-performing schools, all else equal). This may lead to unfair comparisons between schools: are schools with high school performance estimates more effective because these schools generate positive learning effects or because their students are so able and/or socioeconomically advantaged that they would attain the same effects regardless of the school they attended? To account for this kind of bias, researchers and policymakers have used different approaches. Some of these solutions are simple and report school performance without adjusting for students’ backgrounds, so confounding the interrelated role of different determinants. Other approaches use different value-added models with different covariate choices and methodologies to adjust for compositional differences across schools and to disentangle the specific ‘school effect.’

The stability across modeling decisions has been explored in the literature. In the context of teacher value-added estimates, Papay (2011) examines a set of models with a variety of

specifications with respect to student, classroom and school covariates and finds that covariate choice made some difference in ranks of effect estimates, but not as much difference as what he finds when using different testing instruments. The minimum correlation across models exceeds 0.79. Other researchers have reached to similar conclusions when estimating school effects for K-12 (Timmermans et al., 2011) and higher education (Riehl et al., 2016). In contrast, Lockwood, et. al. (2007) find that from 40 to 55 percent of teachers would get noticeably different scores using an alternative value-added model that accounted for student assignments in a different way. A recent development has been proposed by Page et al. (2017) who use the whole distribution of students' test scores within schools to derivate quantile-based measures of schools' value added, in other words estimations about how school effects vary at different point of the distribution of individual students' results. The application to the case of Chilean schools illustrates how important differences emerge when not reducing the performance information at the effect exerted "on average" by the school.

The stability across school performance metrics or outcome choice seems to contribute more to variation in school performance estimates than model specification (Lockwood et al., 2007; Papay, 2011; Riehl et al., 2016). Papay (2011) finds moderate-sized rank correlations, ranging from 0.15 to 0.58, between the estimates derived from different testing instruments or metrics. This instability has been associated with differences in test timing and measurement error across tests. Similarly, but using completely different metrics instead of testing instruments, Riehl et al. (2016) find that correlations between colleges' earnings (as a proxy for long-term outcomes) and learning scores range from -0.09 to 0.71 across different metrics and models.

As a matter of fact, school performance estimates tend to have low stability over time. The literature has looked at the level of stability of school effects finding moderate to low levels of stability (Leckie & Goldstein, 2009; Gorard et al., 2012; Marks, 2014; Allen and Burgess, 2013), and particularly for primary schools (Dumay et al., 2013; Stacy et al., 2013; Tymms & Dean, 2004; Kortelainen et al., 2016). Dumay et al. (2013) explore value-added estimates of different primary grades across time for 1, 2, 3 and 4 years apart, and find relatively low – and declining – correlations of 0.40–0.53, 0.40–0.43, 0.36–0.40 and 0.29 respectively.

Perry (2016) provides several possible explanations for the levels of instability of school effects over time: differences between school performance estimates may not be stable; fluctuations over time in uncontrolled (or unobserved) non-school factors or test scores may provide a noisy measure of school performance due to substantial error variances relative to differences between schools (Kane & Staiger, 2002); or perhaps overcorrecting for compositional differences remove some genuine differences in school effectiveness along with the influences of non-school factors. Taking advantage of a school program assigned to low-performing schools in Chile, Chay et al. (2005) show that transitory testing noise and the consequent mean reversion produce important complications in the evaluation of such programs, suggesting that mean test scores from a single year can provide misleading school ratings.

Instability of school performance estimates over time has been associated with instability across cohorts (Dumay et al., 2013; Marks, 2014). Yet, there are very few studies that have explored the stability of school performance estimates across different cohorts in the same school (Perry, 2016; Teddlie & Reynolds, 2000). An important contribution is due to Thomas et al.

(2007), who analyzed how value-added trends for secondary schools in Great Britain vary in a quite long panel of ten years. The longitudinal perspective highlights how just a very minor proportion of schools were able to maintain a coherent growth trajectory for more than four years, so questioning the institutional and policy ideas behind the concept of “school improvement”. In the perspective of evaluating schools, then, the idea of considering longer time periods is crucial for getting a more reasonable idea of how VA estimates are really measuring changes in performance. Evaluating the stability across cohorts may help to understand if school performance estimates based on a single cohort can predict future performance, or if instability stems from fluctuations in school performance over time or from factors at a given point in time giving rise to inconsistency between cohorts such as differences in teacher effectiveness or teacher quality. This latter eventuality suggests caution in using those indicators of school effects as reliable instruments for formulating predictions about future school performance.

A conceptual issue deserves specific attention, which is rarely highlighted in previous studies. A factor that should be considered as potentially influencing the fluctuations of performance estimations over time and/or across cohorts is the possibility that schools react to unsatisfactory performance at year (t), by adjusting and innovating their activities for improving the performance at years (t+n). For assessing such an eventuality, however, a detailed data collection about school practices would be necessary, and this is very rarely done by evaluation agencies.

Summing up, the existent literature confirms that is possible to identify a school effect and to estimate the value-added by a given school to students’ achievement. This effect must be computed net of those differences across schools that are due to students’ socioeconomic background, which is an element that statistically affects schools’ average performance. The estimated school effects can be in principle used for accountability purposes, as they represent the impact of attending a specific school instead than another. Therefore, many studies highlight that the value-added estimations are quite inconsistent across different modelling choices and over time even for similar specifications.

3. Empirical methodology

A. Data and Sample

De-identified data were provided by the National Evaluation Committee for Education (INVALSI) under a limited-use data agreement¹. INVALSI is the Italian institute responsible for the design, administration and evaluation of annually standardized tests for Italian students. All students enrolled in second, fifth, eighth and tenth grades are required to take two tests, one in mathematics and another one in Italian language. Even though the Italian Ministry of Education emphasizes the importance of establishing a system to evaluate teachers and schools based on students’ performance, the tests have been low stakes for all grades, with the exception of the eighth grade, which corresponds to the end of the compulsory secondary education – in the final exam, the INVALSI test score counts for a (small) portion of the final grade.

¹ We are grateful to the INVALSI’s statistical office for having provided the dataset and particularly to Patrizia Falzetti for the dedicated discussions on earlier versions of the empirical analyses.

From the available data, we requested additional elements from the INVALSI data, including students' socioeconomic characteristics and standardized test scores. We also requested scores from teacher assessments for reading and mathematics formulated during the academic year. The data requested contain test scores and information at student, classroom's and school's level. While the INVALSI tests acquired nationwide status almost a decade ago, for this project we utilize three cohorts of 5th graders data from 2013 through 2015. These cohorts were selected based on the quality of matching between fifth and second grade scores at student level, which significantly improved after 2012. This is an important feature of the dataset because we would include prior achievement in the specification of the educational production function (EPF). The data do not include any measure of teacher or principal characteristics, nor do they include school resources or inputs.

We apply a sample restriction criteria to ensure that all performance measures calculated in this paper are based on the same set of individuals. We restrict our sample to students who satisfy two important criteria. First, we include only fifth grade students who took INVALSI in 2013–2015. Second, we include only students for whom we observe second grade scores. Among key information about scores, there is a statistical approximation about the likelihood that scores have been manipulated at class level (the so called “cheating” phenomenon²), and data used in this paper is corrected for the “propensity to cheating”³.

In addition to these restrictions, we drop individuals with missing values on any of the other key variables we use, including parental education, student SES, measured by the Economic, Social and Cultural Status (ESCS) index,⁴ and other demographic characteristics such as gender, immigration status and student type (i.e. regular, overage, underage). Lastly, and in order to obtain reasonable precision for each of our performance indicators, we restrict our analysis to schools with at least 25 students satisfying the above criteria. This additional restriction partly accounts for the widespread score manipulation on the standardized tests given in primary schools located in the South (Angrist et al., 2016).⁵ In general, our final models include 55 percent of all students in fifth grade and 75 percent of all schools. As shown in ATable 1, the difference in means in demographics and outcome variables between our sample and the population of students in fifth grade who took INVALSI in 2013-2015 are relatively small but statistically different.

² The propensity of cheating was estimated and provided by INVALSI and indicates what portion of the observed score is on average due to school cheating behavior. INVALSI employs cluster analysis, using the following inputs: within-class information on the average and standard deviation of scores, proportion of missing items, and variability in responses. Cluster analysis seeks regularities in data and groups classes with those that are similar to each other. INVALSI identifies classes with manipulated scores as those in an “extreme” cluster. In practice, these are classes with abnormally high performance, small dispersion of scores, low proportion of missing items, and high concentration in response patterns (Battistin, 2016).

³ Using the restrictions described in the text, and by only including observations where the likelihood of cheating was available, we dropped almost 41 percent of observations.

⁴ ESCS is an index for the Economic, Social and Cultural Status of the students' families. This is computed by INVALSI following an OECD's standard, and takes into account parents' occupations and education, along with variables that measure home possession goods (see Campodifiori et al., 2010 for details). It is constructed to obtain a distribution with mean=0 and st.dev=1 in the country.

⁵ Using instrumental variables (IV) estimates, Angrist et al. (2016) show strong class size effects in Southern Italy. The authors suggest that small classes increase manipulation, as a consequence of teacher shirking instead of accountability concerns.

In absolute terms, the resulting sample includes approximately 790,519 of 5th graders from 5,427 schools. For the analyses of consistency using different models and metrics we used the sample presented in Tables 1 and 2. For the analysis of consistency across cohorts, we used the same sample of schools across cohorts, which further restricted our sample to 2,247 schools. The observable characteristics of this sample are mostly statistically insignificant compared to those of our main sample (See ATable 1).

Table 1 and 2 present a summary of statistics for our sample by cohort and school location. Table 1 shows that cohorts of fifth graders have very similar socioeconomic characteristics, while academic characteristics tend to vary by cohort, particularly for average INVALSI fifth-grade scores. At the end of the paper, we explore whether these estimates vary because of the different samples of students who take INVALSI tests in each cohort. Our sample is, however, very heterogeneous when comparing socioeconomic characteristics across regions. Table 2 clarifies the North-South divide in Italy. Students attending schools located in the north are much more likely to have wealthier and more educated parents. For instance, the average SES of students attending schools in the north is around 0.20 while the average SES in the south is negative, -0.50. Another stark evidence from descriptive statistics is that schools in the north tend to perform academically better than those in the south, even when comparing scores from teacher assessments and not only standardized tests.

B. Models and Metrics

This section describes the outcome variables and the process we employ to estimate the resulting metrics for the school effects. After computing regression-adjusted “school fixed effects” to account for compositional differences across schools, we normalize school-level means/fixed effects in order to compare indicators that have different scales, and then assess the resulting metrics using correlation matrices and graphical analysis.

To measure school performance, we construct three learning metrics for the three cohorts of fifth graders. Our learning metrics are based on students’ scores from INVALSI tests and teacher assessments. Both the scores from INVALSI tests and teacher assessments include mathematics and Italian components taken by the students who participated in these tests. Additional details on each metric and its rationale are below.

- 1) *Math and Italian INVALSI Standardized Test Scores (INV)*. We used second and fifth grade INVALSI test scores which include similar but not identical components. INVALSI tests include multiple choice questions and open-response items, for which some grading is required. Students take INVALSI tests in their own schools and are proctored by a teacher from their schools who was not their teacher during the academic year. These teachers are also responsible for grading, transcribing the test scores, and sending them back to INVALSI. Teachers tasked with grading and transcription can enlist colleagues for help. Institutional features that regulate score transcription have been associated to the documented score manipulation, particularly in schools located in the South (Bertoni et al., 2013; Battistin et al., 2014; Pereda Fernandez, 2016).

- 2) *Adjusted Math and Italian INVALSI Standardized Test Scores (WLE)*. The Weighted Likelihood Estimates (WLE) of INVALSI standardized tests include 41 multiple-choice items, summarized by a Rasch score (Andrich, 1988). Rasch builds estimates of true intervals of item difficulty and students' ability by creating linear measures. The purpose of using these weighted scores is to capture any difference in performance that might be explained by the level of difficulty of test items, without relying purely on the proportion of correct answers. Sampling of test items and the distribution of "easy" and "difficult" test items may lead to school or classroom variation in overall test scores. Some schools might have incentives to better prepare students to respond to "easy" test items and perform as good as a school that is more focused on improving overall students' ability. School ratings may then vary when using WLE scores. Nevertheless, the correlation between unadjusted test scores and Rasch measures is very high and highly statistically significant (0.99 for both Math and Italian across models). It is worth noting that INVALSI is currently using these Rasch-adjusted scores to estimate school performance.
- 3) *Math and Italian Teachers' Assessment Scores (TEACH)*. Periodic (quarterly or monthly) and final assessments are taken by teachers in primary and secondary schools in Italy. In daily practice, teachers have the autonomy to determine how to assess students, and student evaluation at all school levels is conducted through oral, written, and practical testing (depending on the subject). We used oral scores in quarterly assessments in mathematics and Italian to complement our analysis.⁶ Because individual-level error from idiosyncratic shocks or classroom-level shocks may cause actual students' INVALSI test scores differ from true scores, it is important to analyze if school effects estimated using INVALSI scores are consistent with other assessments. For instance, some students may be under a lot of stress before INVALSI tests or some teachers may give more time to students to finish the test. While these assessments are taken by the end of the first period of the school year (four months), INVALSI tests are taken in May, towards the end of the school year. Therefore, test timing may also affect school performance estimates. Finally, for school choice, parents might be more interested in knowing how their children perform in school assessments, which are currently higher stake than standardized tests in Italy (i.e. they determine the students' final grades).⁷ It is important to notice that, behind all these correlation studies, the key assumption is that INVALSI tests and teachers' assessment are based on the same set of relevant skills and competences.

For our analysis of stability of school performance across different metrics, we calculate each student's percentile relative to all other students in our sample and cohort. We use score percentiles to be able to control for initial academic ability using second grade INVALSI test scores. Second grade INVALSI scores and scores from teacher assessments are not on a common scale and thus cannot measure growth in human capital in absolute terms. As a result, for this analysis our learning metrics will capture a school's relative rather than absolute performance.

⁶ In our dataset the 81 percent of written scores of teacher assessments for both mathematics and Italian are missing, and 96 percent of the available written scores are equal to the oral ones.

⁷ As anticipated in previous sections, the only notable exception is that of grade 8 (not considered in this paper). Grade 8 coincides with the end of lower secondary school. The final grade is the result of averaging teachers' grades, the scores obtained in a dedicated exam, and INVALSI test scores. The latter component weights 1/7th of the whole final grade.

We use the three metrics of school performance listed above to estimate school effects. The models include different combinations of methods and covariate choice to adjust for compositional differences. Some of the models to estimate school effects are simple and require less-detailed information, and thus they correspond to models that may be more commonly reported in the media or easier for policymakers to compute. Other models use comprehensive information on students' socioeconomic and academic backgrounds and align more closely with "value-added" models currently employed by INVALSI and in the teacher effects literature (e.g. Kane and Staiger 2008, Kane *et al.*, 2013). Once the school fixed effects are estimated, and with the aim to be able to compare across our different indicators, we normalized each of these metrics to mean zero and standard deviation one across the 5,427 schools in our sample. We could also aim at comparing and correlating school ratings, but this is limiting because the ranks eliminate valuable information on how far apart schools are from each other.

We then end up with seven models, which we describe in the following subsections, and that allow us to explore the stability of our results to different modeling decisions. Note that these models control for the propensity of cheating or score manipulation on INVALSI tests, as described in the previous sections. In the Annex, ATable 2 summarizes graphically the covariates included in each of these models.⁸

Computing raw means (Model 1). This method includes a cohort fixed effect but no other covariates, and essentially ranks schools by mean test scores. An obvious concern, however, is that differences in metrics across schools would incorporate essentially many compositional factors other than school performance: they could reflect differences in students' initial ability or students' socioeconomic characteristics (e.g. parental education, immigration status). This suggests the need to adjust these observed means for compositional differences, a process we employ in subsequent analyses.

Computing second grade score residuals (Model 2). Our second performance metric adjusts for differences in initial student academic characteristics by controlling for INVALSI second grade scores (but not for other student-level characteristics). We do this through a student-level regression of the following form:

$$(1) y_{isc} = \beta t_i + \gamma_c + schFE + \varepsilon_{isc}$$

Where y_{isc} is the learning metric for student i who attends school s from cohort c ; t_i is a vector of student i 's score on the INVALSI second grade test and γ_c is cohort fixed effects. We estimate this equation separately for mathematics and Italian. We decompose the residual from this regression into a school-specific term $schFE$ and an idiosyncratic component, ε_{isc} . Our second school performance metric is the coefficient of school fixed effects from equation (1).

Computing second grade scores and SES residuals (Model 3). Our third performance metric includes additional controls for students' socioeconomic background in regression (2):

⁸ All models were also estimated using up to fifth order polynomials of the second grade scores, but for simplicity we settled with linear models because the correlations between these estimates and the linear specifications are very close to 1 (0.99) (See ATable 5).

$$(2) y_{isc} = \beta t_i + \delta X_i + \gamma_c + schFE + \varepsilon_{isc}$$

where x_i represents dummies for student type (i.e. regular, overage, and underage), gender, immigration status, and student SES measured by ESCS student index. This third school performance metric for each school is the coefficient of school fixed effects from this regression. This coefficient is identified from variation in schools' effects across students with the same socioeconomic and academic characteristic. This performance metric is close to "value added" models and emulates previous work (e.g. McCaffrey et al., 2003; Kane and Staiger, 2008). Correcting for compositional differences has been debated in the literature that suggests that overcorrecting may unintentionally remove genuine differences in school performance (Visscher, 2001).

Computing classroom-level residuals (Models 4-6). To control for peer effects and classroom composition, we compute three additional school performance metrics. For the fourth metric, we fit the following model:

$$(3) y_{isc} = \beta t_i + \delta X_i + \partial \bar{X}_j + \gamma_c + schFE + \varepsilon_{isc}$$

where $\bar{X}_j$ is a vector of classroom-level means that include student gender, immigration status, student type, as well as the class size and average second grade scores of the students in the j-th class. The fourth school performance metric is the coefficient of school fixed effects from Equation (3). The fifth metric controls for unobservable and observable classroom characteristics by including classroom fixed effects to Equation (3) in lieu of average classroom characteristics and school fixed effects. This metric for the school effect, here, is calculated as the school average of the coefficient of classroom fixed effects. Finally, to compute the sixth metric we estimate the residuals of Equation (3) without including school fixed effects. We then calculate the average of these residuals for each school.

Computing random effects (Model 7). Lastly, we fit a mixed model with shrunken school effects using Bayes shrinkage estimators and based on the specification described in the Equation (3). It is common in research and policy applications to shrink value-added estimates toward a common Bayesian prior (Herrmann et al., 2013). The benefit of the shrinkage procedure is that it produces estimates for which the estimation-error variance is reduced through the dependence on the stable average school effect. This metric is represented in a multilevel model by a random intercept which is shared by all units (students) belonging to the same school. This random component would account for the residual variability between schools that is not explained by students' characteristics and classroom means. It is worth noting here that this model is similar to what INVALSI is currently using to estimate school performance; indeed, the only difference between Model 7 and INVALSI method is that the latter includes school instead of classroom means in Equation (3). We consider our approach closer to a description of the reality, because it takes into account the potential heterogeneity in performances within schools (i.e. between classes).

Even though different assumptions are necessary when using random effects, Deming (2014) find similar effects when using random or fixed effects, particularly when small classes are excluded from the analytical sample (Jakubowski, 2008). Anticipating our findings, the

correlation of school effects between Model 4 with Model 7, which includes the same covariates but with random instead of fixed effects, are very close to 1 (See Table 5 and Table 7).

4. Empirical Results

A. Consistency across Models

The distribution of the school effects' estimates by model is presented in the density plots of Figures 1 and 2 for mathematics and Italian. Descriptive statistics of school effect estimates are shown in Table 3 and 4. The difference in school performance estimates between top and bottom schools is two standard deviations for both mathematics and Italian and many schools are at most one standard deviation of each other. The distribution of these estimates is very similar across the models employed to adjust for compositional differences across schools. Overall, these statistics reveal that the effects exerted by schools on students' results are quite heterogeneous, and that it is necessary to assess the relative performance of schools, without assuming that their results and influences over students' scores are identical. Nevertheless, we will show later how these differences are often statistically concentrated around "zero effects".

The total variation in school effect has been decomposed to within-school and to between-school variation using Model 7 (see column 5 of Table 10). The total variability is higher for school performance estimates in mathematics than in Italian. The individual-level or within school variance is substantially higher than between-school variance, and therefore we can conclude that only 12 percent of the variation in school performance estimates in mathematics and less than half of this in Italian (4 percent) can be attributed to the schools.⁹

To first explore the role of adjustments for the composition of students' socioeconomic background at school level, Tables 5-8 present, for mathematics and Italian, the correlations between unadjusted and adjusted school effects, where Model 1 corresponds to unadjusted school means and Models 2-7 include student, classroom and school-level covariates. In general, adjusted and unadjusted estimates are very highly positively correlated. Across the seven models employed to estimate school effects, the Pearson and Spearman's rank correlation between the unadjusted effects and adjusted school effects ranges from 0.74 and 0.99 for both Italian and mathematics. Controlling for SES appear to be slightly less important for mathematics than Italian, and classroom observable characteristics seem the least important controls: the correlations between Model 3 and 4 are around 0.98.

Despite the high correlations, some movements in school ratings may influence the judgment of relative school performance, and potentially school choice, particularly when comparing unadjusted with adjusted school effects. To investigate if this is the case, it is necessary to compute how many schools are assigned to a different quartile of the effects' distributions depending upon the different models. Table 9 presents a transition matrix that demonstrates how school choice may play out between Model 1 and Model 7. In general, for both Italian and mathematics, approximately 70 percent of schools will remain in the top or bottom quartile irrespective of compositional adjustments. However, for Italian, 12 percent of schools that rank

⁹ These estimates are the intra-class correlation (ICC) shown in Table 10.

in the bottom quartile using Model 1 would rank in the top half after controlling for compositional factors. This means that schools that seem to perform poorly using mean test scores or unadjusted school effects would perform well when comparing them to schools whose students have similar socio-economic background. A lower percentage of schools (6 percent) report high mean Italian test scores but would underperform when using adjusted school performance estimates. These movements in school ratings are not observed when comparing school performance estimates using adjusted models (See ATable 3), and this can lead to conclude that adjusted models are more robust than unadjusted altogether, as they limit the phenomenon of misattribution of schools to a wrong cluster (quartile). .

To evaluate the role of adjustments in the total variability of performance estimates, we use a different set of covariate choice and fit mixed models with shrunken school effects. As shown in Table 10, without including any covariates results show that most part of the variance is at student-level, even though between-school variance is significantly different from zero, suggesting the existence of some degree of segmentation among schools. Controlling for second grade scores (see column 2 of Table 10) significantly reduce unexplained within and between school variance, confirming the importance of students' initial ability and prior achievement. The inclusion of student demographic characteristics equally reduces the between school variance by 31 percent for both mathematics and Italian, highlighting the importance of compositional effects. Yet, the unexplained variability remains high, stressing the need to check for differences in classroom characteristics in determining differences between schools. However, observable classroom characteristics do not explain much the between-school variance, leaving the question open about the latent factors that are behind the determinants of the estimated school effects.

To determine the extent of the variation in average school performance among schools and whether these estimates are statistically different along the distribution of school ratings, we calculate a 95 percent confidence interval for each school performance estimate using Model 7. Figure 3 and Figure 4 present these results for both Italian and mathematics. Confidence intervals are on average broad, and particularly for the lowest and highest performing schools. This suggests a considerable uncertainty in these estimates at the top and bottom quantiles, although schools that ranked in the middle of the distribution are more indistinguishable from average (i.e. zero), and particularly for Italian. While 82 percent of all schools are statistically indistinguishable from average performance in Italian, 69 percent of schools do not significantly differ from average in mathematics (and around 15 percent of schools are better or worse than average in mathematics compared to 9 percent in Italian).

Finally, our school performance estimates using Model 7 are close to what INVALSI is currently estimating using the latest cohort of fifth graders available. The only methodological difference between Model 7 and INVALSI's model is that instead of using school-level means, we use class means. INVALSI does not report to principals the punctual estimation of the school, but only if their school is in the group of those schools for which the estimated "value added" is negative (lower than the mean), average (not statistically different from the mean) or positive. For the 2015 cohort, as shown in ATable 4, approximately 80 percent of schools in our sample would perform in the same category as what INVALSI is currently reporting to school principals (i.e. positive, average, and negative school performance). In both our models and INVALSI

results, most schools have been deemed to exert an “average” school effect – i.e. indistinguishable from other schools’ average effect.

B. Consistency across cohorts

In order to compare how school performance’s estimates change across cohorts, we differentiate these estimates by cohort. To do this, and as previously explained, we further restrict our sample to re-estimate school effects for the same pool of schools, at different points of time – i.e. for different cohorts.

Firstly, we correlate school performance estimates using all cohorts with a single cohort performance and then assess how persistent school effects are by cohort. Table 11 presents correlations between school-effects’ estimates that use all cohorts versus a single cohort. These correlations are quite high, on average between 0.61 and 0.77 for both Italian and mathematics, meaning that averaging estimates across years is not conducive to results that are drastically different from the ones obtained with a single-year estimation.

Table 12 and Table 13 presents Pearson and Spearman correlations across cohorts for unadjusted and adjusted models. These correlations, instead, are generally low and unstable for both Italian and mathematics, particularly for adjusted models that controls for compositional differences across cohorts. This means that knowing the performance of fifth graders who enter in one year does not accurately predict the performance of the following cohort in the same school – even when adjusting for compositional effects of each cohort. For example, if a school is predicted to be one standard deviation worse than the average school with the 2013 cohort, it is predicted that for the next cohort (the 2014 cohort) this school would be between 0.23 and 0.31 standard deviations worse, for Italian and mathematics respectively. Such substantial differences cast serious doubts about the ability of value added estimations to identify the specific time-persistent effect of attending a given school.

School performance estimates tend to revert to the mean, even after only two years. The correlations of two cohorts apart are lower than for successive cohorts, especially for adjusted models. For instance, school performance estimates from Model 7 predict the next year’s performance on mathematics with a correlation of 0.30 between 2013 and 2014 cohorts, falling by 50 percent to 0.15 when the gap is two years. Our correlations are approximately 50 percent smaller than those found in Perry (2016). He finds an adjusted performance correlation of around 0.5 between two consecutive cohorts, and of 0.24 when comparing performance estimates two years apart at the same school.

Table 14 and Table 15 present transition matrixes using Model 7 for Italian and mathematics, respectively. In general, less than half of schools will remain in the top or bottom quartile irrespective of adjustments. The fact that correlations across cohorts are poorly correlated suggests that each cohort might be influenced differently by other school or student characteristics. Since our dataset does not contain important school characteristics that might influence school choice (e.g. teacher experience, teacher education, principals’ managerial activities and leadership styles, school inputs and resources, etc.), we use student socioeconomic status to explore how single cohort estimates are related to this factor. To do this, we use the

same models described in Section 3 to compute average school performance estimates. Note that we exclude student's SES and school indexes from the adjusted models, as these models include SES controls. Figure 5 shows the correlations of the re-estimated "school effects" with SES for the unadjusted model, Model 1, and Model 7. Even though the estimates of school performance from Model 1 are the most stable, they appear to also be the most confounded with the students' socioeconomic background. For instance, the average correlation between unadjusted performance estimates in Italian for the 2013 cohort and average SES is 0.36. In Model 7, this correlation is lower, 0.29.

In general, the correlations between average SES and school value-added estimates are low and lower for mathematics, ranging from 0.12 to 0.31 for mathematics and from 0.29 to 0.41 for Italian. The performance of schools for the 2015 cohort is slightly less correlated with SES than the other cohorts. Yet, in all cases these correlations are statistically significant at one percent. For Italian, the difference in correlations across cohorts is not statistically significantly different from zero for both adjusted and unadjusted models, suggesting that the low persistence in school performance estimates across cohorts is not influenced by socioeconomic background or compositional differences due to students' characteristics (See Table 16). Mathematics seems to be more influenced by socioeconomic background than the Italian scores, particularly for the latest cohort.

Although our models may not fully isolate "pure" school value added, these findings suggest that the low stability in school performance estimates across cohorts may be due to unobservable characteristics that change over time at school level such as teacher characteristics. In Italy, teachers follow each cohort of students they are assigned to through fifth grade, which means that this factor is a key candidate to explaining the low correlation across different cohorts of fifth graders. The lack of detailed data about teachers' characteristics prevent us to test this hypothesis directly, as well as the absence of a model to assess teachers' quality within and between schools.

C. Correlations across Metrics

For this particular analysis, we use metrics of relative school performance in order to measure growth in human capital by controlling for prior achievement (as measures through the INVALSI second grade scores). Table 17 and 18 shows descriptive statistics of our three metrics for school effects. The differences in school performance estimates between top and bottom schools is two standard deviations for all metrics and across models, indicating that adjustments for compositional differences between schools does not differentially affect school performance estimates across metrics.

Table 19 presents Pearson correlations between school-effect estimates across metrics by models. A simple but important result is that the metrics are mostly positively related to each other, but far from perfectly so, with correlations ranging from -0.01 to 1 for Italian and 0.02 to 1 for mathematics across the three metrics and the seven models. The raw mean school performance estimates (Model 1) are more strongly correlated than the adjusted ones. In fact, as we control for individual and classroom characteristics, the correlations between metrics decrease and become insignificant for both Italian and mathematics. It is also notable that correlations between unadjusted metrics for Italian are slightly higher than for mathematics, but

this pattern is reversed as we control for compositional differences. While INVALSI scores are perfectly correlated with WLE scores (as expected by construction of these two metrics), the correlation of any of these two metrics with scores from teacher assessments is close to zero.

Figure 6 depicts the Spearman correlations between school-effect estimates across metrics for Model 1 and 7. The correlations between INVALSI and WLE scores and the teachers' scores are as low as the Pearson correlations, especially for our adjusted model (Model 7). This means that a school's rank could be dramatically different depending upon the metric used, with clear consequences for the opportunity of developing 'league tables' based on these.

Finally, Table 20 shows the persistence of these metrics across cohorts, indicating that teachers' scores is more stable across cohorts than INVALSI and WLE scores. Also, correlations across cohorts are more sensitive to adjustments for INVALSI and WLE than for teachers' scores, which suggests some consistency in the way students are graded by teachers of the same school across different cohorts of students, even in presence of different results across cohorts in the INVALSI standardized tests.

5. Discussion, implications and concluding remarks

In this paper, we estimate school performance by adopting different models that include different combinations of methods and covariate choice to adjust for compositional differences across schools. Available data include key characteristics of individual students at the end of primary schools (grade 5), such as student SES, gender, age and immigrant status. Moreover, the dataset includes a novel feature in Italy that is the link with performance of students at grade 2 for three consecutive cohorts: this way, the estimations of school effects have a Value Added (VA) interpretation.

Differences in school value-added, although apparent in point estimates, are not substantially large in statistical terms. Most schools are ranked in the middle of the school performance distribution and are indistinguishable from average (i.e. zero), and particularly for Italian. Thus, school ratings are meaningless for driving school choice unless parents only have access to schools ranked at the top or bottom of the school performance distribution.

Even though adjusted and unadjusted estimates are positively correlated, adjustments matter for school choice. Across the seven models employed to estimate school effects, the correlation between the unadjusted effects and adjusted school effects ranges from 0.74 and 0.99 for both Italian and mathematics. Despite the high Pearson and Spearman correlations, movements in school ratings are only observed when comparing school performance estimates using unadjusted models. These findings suggest that school performance estimates are consistent and stable across models even though adjustments affect school ratings in a decisive manner.

At Italian primary schools, school performance estimates are not persistent and correlations across cohorts decline over time. Results suggest that school performance estimates based on a single cohort does not reflect performance at the school more generally; thus, the estimations of school effectiveness are not predictive of future performance. These levels of instability are

consistent with the low to moderate correlations across cohorts from recent research by Perry (2016) and Strand (2016). This instability may be due to fluctuations in unobserved non-school factors or significant measurement error (Gorard et al, 2012). Methods applied to control for compositional differences might also be removing some genuine differences in school performance estimates by reducing bias and making greater demands on the quality of the data (Perry, 2016).

In line with the previous literature (Perry, 2016; Dumay et al., 2013), adjusted school performance correlations across cohorts have lower stability over time than unadjusted correlations. Adjusting for second grade scores and compositional differences seems to leave a residual which is comparatively more unstable than raw scores. This also means that predicting the future school performance based on their past performance is more difficult when schools' performance scores are adjusted by initial compositional differences. Still, even the correlation of unadjusted school performance estimates across cohorts is moderate to low, which suggests that no estimation approach is able to produce a perfect trade-off between bias and stability over time. Therefore, it is crucial to better understand why school performance estimates are not stable and to identify the unobservable school characteristics that determine school effectiveness and change across cohorts. If parents are to choose schools based on performance information when school quality is not persistent, then even unbiased estimates offer little informational value and can distort preferences and choices, so acting in the opposite direction than the one desirable for the policy. Even in more mature contexts of discussion for VA measures – as the UK one - serious doubts have been cast towards the appropriate use of these measure for creating rankings of schools (Wilson & Piebalga, 2008).

In addition, our dataset includes results using two different test instruments, one coming from a standardized test administered at national level, the other reflecting teachers' discretionary scores. The correlation analysis of the 'school effects' using a difference choice of metric indicate major differences in school ratings. These differences might stem from differences in test content and timing, scaling, and measurement error, with teachers' scores being on average lower in Italian and in general slightly more stable over time than standardized scores.

The key findings presented here, along with the lessons drawn from them, are of interest for international readers, because the main problems arising from the implementation over time of an accountability system for education are the same in any institutional context. As a consequence, the methodological challenges discussed in this paper for the Italian case are analogous to those already faced in countries where accountability systems have been implanted several years ago, like for example UK, USA and Sweden (see the discussion of the British case in Leckie & Goldstein, 2017, and that of US experience in Collins & Amrein-Beardsley, 2014). The present research aims then at extending the debate about the risks and opportunities related with an expansion of policies based on accountability and school choice; in so doing, we focus in particular on the credibility of school performance's estimates, as potentially developed by the Ministry of Education and other regulators, policy makers or media producers.

In a series of related paper, multilevel statistical methods have been applied to INVALSI data on middle schools (grades 6-8), and revealed that middle schools are characterized by heterogeneity of performance (Agasisti et al., 2017a; b), while this is much less the case in the

primary schools according to the findings presented in this paper. Acknowledging these differences in schools' performance may be important in an accountability perspective, also with the aim of informing better choices by parents. At the same time, the inconsistency of school effects' estimates across the metrics, specifications and cohorts suggests a very cautious approach in deciding how using results about using schools' value-added. At this stage, INVALSI calculates these effects and communicates them to school principals, but not to the public. Moreover, according to INVALSI's analyses, the school effects for the majority of schools is not statistically different from zero. Thus, messages that can lead to school ratings – without considering how similar their value added actually is – might be misleading of actual performance differentials.

The low correlation between school effects estimates using standardized tests vs teachers' scores suggests the necessity of investigating the nature of these differences more profoundly. Of course, there are reasonable arguments to justify of these differences, for example (i) the different contents of tests and topics evaluated by the teachers, and/or (ii) the natural propensity of teachers to grade students on the basis of multiple tests and periods of observation while the test analyses the performance in a single day. On the other side, it can be the case that standardized tests contribute to emerging differences in skills and competences which are not adequately graded by teachers. Moreover, the teachers grade in a discretionary way by construction, thus they can assign different values and weights to specific contents and skills when compared to purely standardized tests. Such a discretion can be a negative factor for accountability, to the extent that they mask genuine differences in schools' performance. In Italy, there is discussion among experts and commentators about the potential negative effects of using only discretionary grades for accountability purposes. For instance, the average grades assigned by teachers at the end of secondary education (grade 13) are much higher in the schools located in the South than in the North, while the evidence from all the standardized tests (not only INVALSI, but also OECD's PISA and IEA's TIMSS and PIRLS¹⁰) indicate that students attending schools in the North have better measured skills than their counterparts in the South.

The estimates of school effects based on teachers' scores are nevertheless more consistent over time and student cohorts than those obtained through standardized tests. This higher robustness over time is likely to be determined by the work conducted within schools to align and coordinate the evaluation criteria between teachers. When this work is conducted, a potential consequence is that the teachers assign grades on the basis of a coordination between colleagues for the same cohort of students, and more homogeneity is guaranteed over time. Conversely, such an approach can result in distorting estimations of the school effect on standardized tests, if the different cohorts of students have different levels of ability, engagement and motivation. Generally speaking, it seems more useful to complement the discretionary judgments of teachers with more "objective" measures derived from standardized tests, with the awareness that neither the former nor the latter can be used as perfect indications of students and schools' performances. In the attempt of enlarging the set of information available for parents' choices, the disclosure of data about school performance coming from the two different approaches

¹⁰ The acronyms refer to the Programme for International Student Assessment (PISA) conducted by OECD and to the two assessments administered by IEA: Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS).

(teachers and standardized tests) can guarantee a pluralistic and multidimensional view on the institutions' results.

In the context of the uncertainty about the magnitude and direction of school effects on students' performance, it is important to reflect on the estimation of the likely impacts of specific policies and educational intervention under discussion in Italy. For example, class reductions have been found to have no effect on student learning (Angrist et al., 2017); also, Ballatore et al. (2014) show an excessive concentration of immigrant students in a given class can be detrimental for the learning of the whole group of students attending that class. The findings presented in this research suggest that some cohort-specific elements can be at work, and it can be difficult to disentangle the effects of these specificities from the consequences of the policies and interventions proposed. In this perspective, improving the methods for school evaluation can go hand-in-hand with the necessity of developing new and better approaches for policy evaluation.

This paper paves the way for further research to be conducted in the next future. In particular, we believe that three areas deserve more attention from researchers in the area.

First, here we do not explore the North-South gap in any way. Previous empirical studies revealed a huge gap in the performance of students and schools in the different geographical areas of the country (Bratti et al., 2007); in addition, preliminary results using INVALSI data indicate that the gap between North and South is not only about the absolute test scores, but also deal with the impact of schools on students' performance (see Agasisti et al., 2017b). Undergoing studies are applying the methods and perspectives presented in this paper in a differentiated way across geographical areas of Italy, with the aim of understanding whether the findings that we obtained (i.e. inconsistency of school effect estimates) hold homogeneously across regions.

Second, this paper does not explore the role of school management and resources in affecting the performance of each institution. Some recent literature suggests that school management does matter in influencing the achievement levels of students (see Bloom et al., 2015; Di Liberto et al., 2015). Explorations about such relationships have been conducted also in Italy, and the empirical results demonstrate that certain managerial practices can be conducive to better performance of the schools (Agasisti et al., 2016). In the current research, we do not have data about resources invested by each school and/or about the characteristics of school principals and their practices and activities. Such lack of data limits the ability of identifying the main drivers of school effects. A more systematic and complete collection of administrative data at school level could represent a major improvement in estimations of school performance, which are at the basis of any accountability reform.

Third, we do not have data about teachers. Reading throughout all the findings, the analyst understands that much of the instability of school effects' estimates are due to time-variant differences between schools – and the quality of teachers who follow the different cohort of students is the primary reasonable explanation. In a different context and with different aims, some evidence about the role of teachers in the production of educational results in Italy has been provided by Abbiati et al (2017). The authors show how teachers with less desirable characteristics (lower formal education and less experience) are more often matched with classes

where the students have a lower socioeconomic background. Their focus is on variation within schools, and they point at explaining a substantial part of the variation in students' results within school. Nevertheless, an identical conceptual framework could be applied to differences between schools and use the variation in teachers' features to characterize interesting patterns in the estimated "school-effects". Lack of data impedes this development. For the next future, it would be important that the Ministry of Education, in its effort of improving the datasets available for research and policy-making, puts a particular more attention to this area of teachers' characteristics. Indeed, this enrichment of the database would represent another fundamental step to understand the real factors associated with differences between the performance of schools, and subsequent policies could be proposed and implement.

References

- Abbiati G., Argentin G., & Gerosa T.. (2017). Different teachers for different students? Evidence on teacher-student matching and its consequences in the Italian case. *Politica economica*, (1), 12-58.
- Agasisti, T., Falzetti, P., & Soncin, M. (2016). Italian school principals' managerial behaviors and students' test scores: an empirical analysis. Working Paper of the Italian Society for Public Economics, No. 43.
- Agasisti, T., Ieva, F., & Paganoni, A. M. (2017b). Heterogeneity, school-effects and the North/South achievement gap in Italian secondary education: evidence from a three-level mixed model. *Statistical Methods & Applications*, 26(1), 157-180.
- Agasisti, T., Masci, C., Ieva, F., & Paganoni, A. M. (2017a). Bivariate multilevel models for the analysis of mathematics and reading pupils' achievements. *Journal of Applied Statistics*, 44(7), 1296-1317.
- Allen, R. & Burgess, S. (2013) Evaluating the provision of school performance information for school choice, *Economics of Education Review*, 34, 175–190.
- Angrist, J. D., Battistin, E., & Vuri, D. (2017). In a small moment: class size and moral hazard in the Italian Mezzogiorno, *American Economic Review*, forthcoming.
- Ballatore, R., Fort, M., & Ichino, A. (2014). Babel's tower in the classroom? Immigrants and natives in Italian schools. In IZA Discussion Paper No. 8732.
- Bertoni, M., G. Brunello, & L. Rocco (2013): When the Cat Is Near, the Mice Won't Play: The Effect of External Examiners in Italian Schools, *Journal of Public Economics*, 104, 65–77.
- Black, S. E., & Machin S. (2011). Housing Valuation and School Performance. Chap. 10 in *Handbook in Economics of Education*, vol. 3, edited by Eric Hanushek, Stephen Machin, and Ludger Woessmann. Elsevier.
- Black, S. E. (1999). Do Better Schools Matter? Parental Valuation of Elementary Education. *Quarterly Journal of Economics* 114(2): 577-99.
- Bloom, N., Lemos, R., Sadun, R., & Van Reenen, J. (2015). Does management matter in schools?. *The Economic Journal*, 125(584), 647-674.
- Bratti, M., Checchi, D., & Filippin, A. (2007). Geographical differences in italian students' mathematical competencies: evidence from PISA 2003. *Giornale degli Economisti e Annali di Economia*, 299-333.
- Burgess, S., Greaves, E., Vignoles, A., & Wilson, D. (2015). What Parents Want: School Preferences and School Choice. *The Economic Journal*, Volume 125, Issue 587.
- Campodifiori, E., Figura, E., Papini, M., & Ricci, R. (2010). Un indicatore di status socio-economico-culturale degli allievi della quinta primaria in Italia. Retrieved March, 5, 2011.
- Chay, Kenneth Y., Patrick J. McEwan, & Miguel Urquiola. (2005). The Central Role of Noise in Evaluating Interventions That Use Test Score to Rank Schools. *American Economic Review* 95(4): 1237-58.
- Collins, C., & Amrein-Beardsley, A. (2014). Putting growth and value-added models on the map: A national overview. *Teachers College Record*, 116(1), 1-32.

- Deming, D. J., & Figlio, D. (2016). Accountability in US education: Applying lessons from K-12 experience to higher education. *Journal of Economic Perspectives*, 30(3), 33-55. DOI: 10.1257/jep.30.3.33.
- Deming, D. J. (2014). Using School Choice Lotteries to Test Measures of School Effectiveness. *American Economic Review* 104(5): 406-411.
- Di Liberto, A., Schivardi, F., & Sulis, G. (2015). Managerial practices and student performance. *Economic Policy*, 30(84), 683-728.
- Dumay, X., Coe, R. & Anumendem, D.N. (2013) Stability over time of different methods of estimating school performance, *School Effectiveness and School Improvement*, 25(1), 64–82.
- Everson, K. C. (2017). Value-Added Modeling and Educational Accountability: Are We Answering the Real Questions?. *Review of Educational Research*, 87(1), 35-70.
- Figlio, David N., & Kenny, Lawrence W. (2009). Public Sector Performance Measurement and Stakeholder Support. *Journal of Public Economics* 93(9-10): 1069-77
- Figlio, David N., & Lucas M. (2004). What's in a Grade? School Report Cards and the Housing Market. *American Economic Review* 94(3): 591-604.
- Figlio, David N., & Rouse, C. (2006). Do Accountability and Voucher Threats Improve Low-Performing Schools? *Journal of Public Economics* 90 (1-2): 239-55
- Gorard, S. (2012) The propagation of errors in experimental data analysis: A comparison of pre- and post-test designs, *International Journal of Research & Method in Education*, 36(4), 1–14.
- Hastings, Justine S., & Weinstein, Jeffrey M. (2008). Information, School Choice, and Academic Achievement: Evidence from Two Experiments. *Quarterly Journal of Economics* 123 (4) : 1373-1414.
- Herrmann, M., Elias Walsh, Eric Isenberg & Alexandra Resch. 2013. Shrinkage of Value Added Estimates and Characteristics of Students with Hard-to-Predict Achievement Levels. Policy Report, Mathematica Policy Research.
- Holbein J. & Ladd H. (2017). Accountability pressure: Regression discontinuity estimates of how No Child Left Behind influenced student behavior. *Economics of Education Review* 58 (2017) 55-67.
- Jacob, B. A. (2002). Where the boys aren't: Non-cognitive skills, returns to school and the gender gap in higher education. *Economics of Education review*, 21 (6), 589–598.
- Jakubowski, M. (2008). Implementing value-added models of school assessment (EUIRSCAS Working Paper No. 2008/06). Florence, Italy: European University Institute. Retrieved from <http://ideas.repec.org/p/rsc/rsceui/2008-06.html>
- Kane, T.J., & Staiger, D.O.(2008). Estimating teacher impacts on student achievement: An experimental evaluation (NBER Working Paper No.14607).
- Kane, Thomas J., & Staiger, Douglas O. (2002). The Promise and Pitfalls of Using Imprecise School Accountability Measures. *Journal of Economic Perspectives* 16(4): 91-114
- Kortelainen, M., Pursiainen, H., & Pääkkönen, J. 2016. "Value-Added Rankings of Finnish Upper-Secondary Schools." [In Finnish] VATT Research Reports 179.
- Leckie, G. & Goldstein, H. (2009) The limitations of using school league tables to inform schoolchoice, *Journal of the Royal Statistical Society*, 172, 835–851.

- Leckie, G., & Goldstein, H. (2017). The evolution of school league tables in England 1992–2016: ‘Contextual value-added’, ‘expected progress’ and ‘progress 8’. *British Educational Research Journal*, 43(2), 193-212.
- Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, 44(1), 47–67. London, UK: Falmer Press.
- Marks, G.N. (2014) The size, stability, and consistency of school effects: Evidence from Victoria, *School Effectiveness and School Improvement*, 26(3), 397–414.
- McCaffrey, Daniel F., Lockwood, J. R., Koretz, D., Louis, Thomas A. & Laura Hamilton, 2004. "Models for Value-Added Modeling of Teacher Effects," *Journal of Educational and Behavioral Statistics*, , vol. 29(1), pages 67-101, March.
- Ministry of Education, Dichiarazioni programmatiche, 2001 - <http://www.istruzione.it>
- OECD (2012). Programme for International Student Assessment (PISA) Results from PISA 2012. (Paris, Organization for Economic Cooperation and Development).
- OECD (2016). Programme for International Student Assessment (PISA) Results from PISA 2015 - Excellence and Equity in Education. (Paris, Organization for Economic Cooperation and Development).
- Page, G. L., Martín, E. S., Orellana, J., & González, J. (2017). Exploring complete school effectiveness via quantile value added. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(1), 315-340.
- Papay, J. P. (2011). Different tests, different answers: The stability of teacher value added estimates across outcome measures. *American Educational Research Journal*, 48, 163–193. doi:10.3102/0002831210362589
- Perry, T. (2016). English Value-Added Measures: Examining the Limitations of School Performance Measurement. *British Educational Research Journal* Vol. 42, No. 6, December 2016, pp. 1056–1080 DOI: 10.1002/berj.3247
- Riehl, E., & Saavedra, Juan E. & Urquiola, M. (2017). Learning and Earning: An Approximation to College Value Added in Two Dimensions, NBER Chapters, in: *Productivity in Higher Education* National Bureau of Economic Research, Inc.
- Santiago Pereda Fernández. (2016). A new method for the correction of test scores manipulation. *Temì di discussione (Economic working papers)* 1047, Bank of Italy, Economic Research and International Relations Area.
- Stacy, Brian & Guarino, Cassandra M. & Reckase, Mark D. and Wooldridge, Jeff. (2013) Does the Precision and Stability of Value-Added Estimates of Teacher Performance Depend on the Types of Students They Serve?. IZA Discussion Paper No. 7676. Available at SSRN: <https://ssrn.com/abstract=2345593>
- Strand, S. (2016) Do some schools narrow the gap? Differential school effectiveness revisited, *Review of Education*, 4(2), 107–144.
- Teddlie, C., & Reynolds, D. (2000). The international handbook of school effectiveness research.
- Thomas, S., Peng, W. J., & Gray, J. (2007). Modelling patterns of improvement over time: value added trends in English secondary school performance across ten cohorts. *Oxford Review of Education*, 33(3), 261-295.

- Timmermans, A. C., Doolaard, S., & de Wolf, I. (2011). Conceptual and empirical differences among various value-added models for accountability. *School Effectiveness and School Improvement*, 22(4), 393-413.
- Tymms, P. & Dean, C. (2004) Value-added in the primary school league tables: A report for the National Association of Head Teachers (Durham, Curriculum, Evaluation Management Centre).
- Visscher, A.J. (2001) Public school performance indicators: Problems and recommendations, *Studies in Educational Evaluation*, 27(3), 199–214.
- Wilson, D., & Piebalga, A. (2008). Performance measures, ranking and parental choice: an analysis of the English school league tables. *International Public Management Journal*, 11(3), 344-366.

Table 1. Descriptive Statistics by Cohort of 5th Graders

Variables	2013	2014	2015	Total
<i>Demographics and SES</i>				
Female	0.50	0.50	0.50	0.50
Italian	0.97	0.97	0.97	0.97
Regular Student	0.98	0.98	0.98	0.98
Average Student SES	0.14	0.10	0.11	0.11
Father in High Income Occupation	0.22	0.23	0.22	0.22
Mother in High Income Occupation	0.11	0.12	0.12	0.11
Parents Attended College	0.67	0.70	0.71	0.69
<i>Outcome Metrics</i>				
INVALSI Score in Ma, 5th Grade	65.26	58.62	56.50	59.71
INVALSI Score in Ma, 2nd Grade	60.47	60.91	58.76	59.86
INVALSI Score in It, 5th Grade	63.40	59.80	65.65	63.38
INVALSI Score in It, 2nd Grade	68.32	69.85	62.28	66.16
WLE in Math, 5th Grade	204.68	206.39	206.18	205.79
WLE in Italian, 5th Grade	203.12	205.91	203.12	203.88
Teacher's Score in Math, 5th Grade	7.92	7.94	7.96	7.94
Teacher's Score in Italian, 5th Grade	7.84	7.87	7.90	7.87
Number of Students	237,526	215,533	337,460	790,519
Number of Schools	3,706	3,256	4,929	5,427

Note: Average INVALSI and WLE scores are corrected by the propensity to test score manipulation using INVALSI estimations. Students' SES is measured by the ESCS index. The ESCS is an index of economic, social and cultural status that was provided by INVALSI and was estimated using availability of household assets, parental education and employment status. We used oral scores for teacher assessments because written examinations had 80 percent of missing values. Our sample includes only fifth grade students who took INVALSI in 2013–2015, for whom we observe second grade scores. We exclude students with missing values on any of the other variables we use as covariates. Lastly, we restrict our analysis to schools with at least 25 students satisfying the above criteria.

Table 2. School Characteristics by Region

School Location	North West	North East	Center	South	Island	Total
INV. Score in Ma, 5th Grade	59.74	60.25	59.84	58.24	57.22	59.15
INV. Score in Ma, 2nd Grade	60.92	60.81	60.93	56.88	57.10	59.36
INV. Score in it, 5th Grade	64.47	64.06	64.58	61.44	61.34	63.23
INV. Score in it, 2nd Grade	67.82	67.80	66.19	61.66	60.88	65.05
WLE in Math	207.15	207.96	206.77	201.84	199.44	204.89
WLE in Italian	206.94	206.54	205.95	196.96	194.66	202.56
Teacher's Score in Math, 5th Grade	8.00	8.01	7.93	7.81	7.93	7.93
Teacher's Score in Italian, 5th Grade	7.92	7.93	7.88	7.76	7.88	7.87
Avg. Student SES	0.19	0.22	0.26	-0.05	-0.11	0.11
Father in High Inc. Occupation	0.24	0.24	0.25	0.21	0.19	0.23
Mother in High Inc. Occupation	0.14	0.13	0.13	0.09	0.08	0.11
Parents Attended College	0.71	0.72	0.77	0.65	0.62	0.69
No. of Schools	1,405	949	969	1,486	618	5,427
If the School is Public	0.88	0.92	0.93	0.95	0.96	0.92
No. of Students	240,168	183,564	126,066	188,104	52,617	790,519

Note: Average INVALSI and WLE scores are corrected by the propensity to test score manipulation using INVALSI estimations. Students' SES is measured by the ESCS index. The ESCS is an index of economic, social and cultural status that was provided by INVALSI and was estimated using availability of household assets, parental education and employment status. We used oral scores for teacher assessments because written examinations had 80 percent of missing values. Our sample includes only fifth grade students who took INVALSI in 2013–2015, for whom we observe second grade scores. We exclude students with missing values on any of the other variables we use as covariates. Lastly, we restrict our analysis to schools with at least 25 students satisfying the above criteria. Average estimates are reported for three cohorts of fifth graders.

Table 3. Descriptive Statistics of school-effects' estimates across Models-Italian

Model	Mean	1st Decile	1st Quartile	3rd Quartile	9th Decile
M1	0.042	-0.961	-0.431	0.553	1.028
M2	0.069	-0.853	-0.334	0.534	0.958
M3	0.068	-0.834	-0.328	0.499	0.958
M4	0.063	-0.852	-0.368	0.507	0.976
M5	0.072	-0.830	-0.318	0.502	0.958
M6	0.056	-0.856	-0.373	0.495	0.961
M7	0.063	-0.931	-0.417	0.550	1.048

Note: Our sample includes only fifth grade students who took Italian INVALSI tests in 2013–2015, for whom we observe second grade scores. We exclude students with missing values on any of the other variables we use as covariates. Lastly, we restrict our analysis to schools with at least 25 students satisfying the above criteria. Our sample is further restricted to schools with at least 25 students satisfying the above criteria (5,427 schools). These estimates use weights equal to each school's number of students.

Table 4. Descriptive Statistics of school-effects' estimates across Models-Mathematics

Model	Mean	1st Decile	1st Quartile	3rd Quartile	9th Decile
M1	0.021	-0.971	-0.478	0.516	1.022
M2	0.070	-0.899	-0.330	0.545	0.926
M3	0.067	-0.897	-0.355	0.529	0.929
M4	0.050	-0.912	-0.411	0.504	0.995
M5	0.069	-0.885	-0.346	0.529	0.920
M6	0.050	-0.896	-0.402	0.505	0.971
M7	0.051	-0.957	-0.441	0.537	1.054

Note: Our sample includes only fifth grade students who took Math INVALSI tests in 2013–2015, for whom we observe second grade scores. We exclude students with missing values on any of the other variables we use as covariates. Lastly, we restrict our analysis to schools with at least 25 students satisfying the above criteria. Our sample is further restricted to schools with at least 25 students satisfying the above criteria (5,427 schools). These estimates use weights equal to each school's number of students.

Table 5. Pearson Correlations between school-effects' estimates across Models-Italian

	M1	M2	M3	M4	M5	M6	M7
M1	1.000	0.000					
M2	0.807 *	1.000					
M3	0.758 *	0.961 *	1.000				
M4	0.735 *	0.937 *	0.981 *	1.000			
M5	0.801 *	0.899 *	0.974 *	0.957 *	1.000		
M6	0.744 *	0.965 *	0.999 *	0.979 *	0.964 *	1.000	
M7	0.795 *	0.913 *	0.980 *	0.962 *	0.996 *	0.974 *	1.000

Note: Our sample includes only fifth grade students who took INVALSI tests in 2013–2015, for whom we observe second grade scores. We exclude students with missing values on any of the other variables we use as covariates. Lastly, we restrict our analysis to schools with at least 25 students satisfying the above criteria. Our sample is further restricted to schools with at least 25 students satisfying the above criteria (5,427 schools). Models are further described in ATable 2. These correlations use weights equal to each school's number of students. Stars indicate statistical significance at 1 percent.

Table 6. Spearman Correlations between school-effects' estimates across Models-Italian

	M1	M2	M3	M4	M5	M6	M7
M1	1.000	0.000					
M2	0.801 *	1.000					
M3	0.745 *	0.950 *	1.000				
M4	0.721 *	0.922 *	0.978 *	1.000			
M5	0.779 *	0.879 *	0.970 *	0.952 *	1.000		
M6	0.732 *	0.955 *	0.998 *	0.976 *	0.960 *	1.000	
M7	0.776 *	0.896 *	0.978 *	0.958 *	0.996 *	0.971 *	1.000

Note: Our sample includes only fifth grade students who took INVALSI tests in 2013–2015, for whom we observe second grade scores. We exclude students with missing values on any of the other variables we use as covariates. Lastly, we restrict our analysis to schools with at least 25 students satisfying the above criteria. Our sample is further restricted to schools with at least 25 students satisfying the above criteria (5,427 schools). Models are further described in ATable 2. Stars indicate statistical significance at 1 percent.

Table 7. Pearson Correlations between school-effects' estimates across Models-Mathematics

	M1		M2		M3		M4		M5		M6		M7
M1	1.000												
M2	0.762	*	1.000										
M3	0.754	*	0.979	*	1.000								
M4	0.741	*	0.961	*	0.986	*	1.000						
M5	0.840	*	0.905	*	0.960	*	0.951	*	1.000				
M6	0.734	*	0.982	*	0.999	*	0.985	*	0.950	*	1.000		
M7	0.844	*	0.935	*	0.973	*	0.962	*	0.993	*	0.965	*	1.000

Note: Our sample includes only fifth grade students who took INVALSI tests in 2013–2015, for whom we observe second grade scores. We exclude students with missing values on any of the other variables we use as covariates. Lastly, we restrict our analysis to schools with at least 25 students satisfying the above criteria. Our sample is further restricted to schools with at least 25 students satisfying the above criteria (5,427 schools). Models are further described in ATable 2. These correlations use weights equal to each school's number of students. Stars indicate statistical significance at 1 percent.

Table 8. Spearman Correlations between school-effects' estimates across Models-Mathematics

	M1		M2		M3		M4		M5		M6		M7
M1	1.000												
M2	0.766	*	1.000										
M3	0.753	*	0.975	*	1.000								
M4	0.740	*	0.956	*	0.985	*	1.000						
M5	0.831	*	0.898	*	0.958	*	0.948	*	1.000				
M6	0.734	*	0.978	*	0.999	*	0.984	*	0.947	*	1.000		
M7	0.838	*	0.929	*	0.971	*	0.959	*	0.993	*	0.963	*	1.000

Note: Our sample includes only fifth grade students who took INVALSI tests in 2013–2015, for whom we observe second grade scores. We exclude students with missing values on any of the other variables we use as covariates. Lastly, we restrict our analysis to schools with at least 25 students satisfying the above criteria. Our sample is further restricted to schools with at least 25 students satisfying the above criteria (5,427 schools). Models are further described in ATable 2. Stars indicate statistical significance at 1 percent.

Table 9. Transition Matrices of School Quartiles (based on school-effects' estimates) for Italian (top panel) and Mathematics (bottom panel)

Italian: Model 7 vs. Model 1				
M1 Quartiles	M7 Quartiles			
	Top	Q2	Q3	Bottom
Top	71%	23%	5%	1%
Q2	20%	46%	28%	7%
Q3	6%	22%	45%	26%
Bottom	3%	9%	22%	66%

Mathematics: Model 7 vs. Model 1				
M1 Quartiles	M7 Quartiles			
	Top	Q2	Q3	Bottom
Top	73%	23%	3%	0%
Q2	20%	48%	29%	3%
Q3	6%	23%	48%	23%
Bottom	1%	6%	20%	73%

Note: Estimates represent unadjusted raw means from Model 1 and school random effects from Model 7. Our sample includes only fifth grade students who took INVALSI in 2013–2015, for whom we observe second grade scores. We exclude students with missing values on any of the other variables we use as covariates. Our sample is further restricted to schools with at least 25 students satisfying the above criteria (5,427 schools).

Table 10. Variance of school-effects' estimates, considering Model 7

Random Effects				
	(1)	(2)	(3)	(4)
Italian	No Controls	Student's Second Grade Score	Student Demographic Control	Class Means
Between variance	33.88	22.06	15.12	14.57
Within variance	273.57	198.13	185.06	183.39
Total variance	307.46	220.19	200.18	197.96
ICC	11.0%	10.0%	7.6%	7.4%
Change in between variance		-34.9%	-31.5%	-3.6%
Change in within variance		-27.6%	-6.6%	-0.9%

Random Effects				
	(1)	(2)	(3)	(4)
Mathematics	No Controls	Student's Second Grade Score	Student Demographic Control	Class Means
Between variance	60.65	39.70	27.23	26.43
Within variance	316.50	223.70	205.45	202.09
Total variance	377.14	263.40	232.68	228.52
ICC	16.1%	15.1%	11.7%	11.6%
Change in between variance		-34.5%	-31.4%	-2.9%
Change in within variance		-29.3%	-8.2%	-1.6%

Note: This Table presents the decomposition of the total variation in school effect. For these estimates we fitted a mixed model with shrunken school effects using Bayes shrinkage estimators Estimates in Column 4 correspond to Model 7.

Table 11. Pearson Correlation between school-effects' estimates using All Cohorts Versus a Single Cohort by Model

		M1		M2		M3		M4		M5		M6		M7	
<i>Italian</i>															
	2013	0.772	*	0.665	*	0.649	*	0.672	*	0.623	*	0.663	*	0.673	*
	2014	0.763	*	0.700	*	0.674	*	0.678	*	0.686	*	0.676	*	0.681	*
	2015	0.782	*	0.679	*	0.654	*	0.658	*	0.349	*	0.666	*	0.662	*
<i>Mathematics</i>															
	2013	0.707	*	0.646	*	0.627	*	0.628	*	0.604	*	0.623	*	0.629	*
	2014	0.773	*	0.724	*	0.711	*	0.711	*	0.677	*	0.712	*	0.714	*
	2015	0.781	*	0.677	*	0.685	*	0.710	*	0.561	*	0.705	*	0.710	*

Note: This table displays Pearson correlations between school performance estimates controlling by cohort fixed effects with estimates for each cohort separately. Our sample includes only fifth grade students who took INVALSI in 2013–2015, for whom we observe second grade scores. We exclude students with missing values on any of the other variables we use as covariates. Our sample is further restricted to schools with at least 25 students satisfying the above criteria and to the same pool of schools across cohorts (2,247 schools). These correlations use weights equal to each school's number of students. Stars indicate statistical significance at 1 percent.

Table 12. Pearson Correlation between school-effects' estimates across Cohorts by Model

	M1		M2		M3		M4		M5		M6		M7	
<i>Italian</i>														
Corr(2013, 2014)	0.424	*	0.242	*	0.202	*	0.226	*	0.273	*	0.223	*	0.231	*
Corr(2014, 2015)	0.432	*	0.293	*	0.247	*	0.257	*	0.000		0.260	*	0.264	*
Corr(2013, 2015)	0.404	*	0.149	*	0.110	*	0.146	*	-0.079	*	0.139	*	0.147	*
<i>Mathematics</i>														
Corr(2013, 2014)	0.424	*	0.346	*	0.317	*	0.307	*	0.308	*	0.289	*	0.306	*
Corr(2014, 2015)	0.368	*	0.202	*	0.194	*	0.226	*	0.170	*	0.227	*	0.226	*
Corr(2013, 2015)	0.322	*	0.119	*	0.110	*	0.152	*	0.109	*	0.150	*	0.152	*

Note: This table displays Pearson correlations for three cohorts in the sample school by Model. Our sample includes only fifth grade students who took INVALSI in 2013–2015, for whom we observe second grade scores. We exclude students with missing values on any of the other variables we use as covariates. Our sample is further restricted to schools with at least 25 students satisfying the above criteria and to the same pool of schools across cohorts (2,247 schools). These correlations use weights equal to each school's number of students. Stars indicate statistical significance at 1 percent.

Table 13. Spearman Correlation between school-effects' estimates across Cohorts by Model

	M1		M2		M3		M4		M5		M6		M7	
<i>Italian</i>														
Corr(2013, 2014)	0.402	*	0.247	*	0.211	*	0.237	*	0.255	*	0.237	*	0.241	*
Corr(2014, 2015)	0.388	*	0.226	*	0.186	*	0.214	*	0.061	*	0.214	*	0.218	*
Corr(2013, 2015)	0.375	*	0.146	*	0.105	*	0.141	*	0.000		0.134	*	0.143	*
<i>Mathematics</i>														
Corr(2013, 2014)	0.403	*	0.302	*	0.282	*	0.297	*	0.278	*	0.286	*	0.298	*
Corr(2014, 2015)	0.356	*	0.195	*	0.196	*	0.230	*	0.185	*	0.223	*	0.229	*
Corr(2013, 2015)	0.297	*	0.116	*	0.113	*	0.159	*	0.106	*	0.148	*	0.158	*

Note: This table displays Spearman correlations for three cohorts in the sample school by Model. Our sample includes only fifth grade students who took INVALSI in 2013–2015, for whom we observe second grade scores. We exclude students with missing values on any of the other variables we use as covariates. Our sample is further restricted to schools with at least 25 students satisfying the above criteria and to the same pool of schools across cohorts (2,247 schools). Stars indicate statistical significance at 1 percent.

Table 14. Transition Matrices of School Quartiles (based on school-effects' estimates) by Cohort – Italian

Italian: 2013 vs. 2014				
2013 Quartiles	2014 Quartiles			
	Top	Q2	Q3	Bottom
Top	40%	25%	19%	16%
Q2	24%	28%	27%	21%
Q3	18%	27%	28%	27%
Bottom	18%	20%	25%	36%

Italian: 2014 vs. 2015				
2014 Quartiles	2015 Quartiles			
	Top	Q2	Q3	Bottom
Top	37%	25%	19%	19%
Q2	25%	29%	25%	21%
Q3	20%	25%	30%	26%
Bottom	18%	22%	26%	34%

Italian: 2013 vs. 2015				
2013 Quartiles	2015 Quartiles			
	Top	Q2	Q3	Bottom
Top	36%	25%	20%	19%
Q2	20%	29%	28%	23%
Q3	21%	24%	29%	27%
Bottom	22%	23%	24%	31%

Note: Estimates represent school random effects from Model 7. Our sample includes only fifth grade students who took INVALSI in 2013–2015, for whom we observe second grade scores. We exclude students with missing values on any of the other variables we use as covariates. Our sample is further restricted to schools with at least 25 students satisfying the above criteria and to the same pool of schools across cohorts (2,247 schools).

Table 15. Transition Matrices of School Quartiles (based on school-effects' estimates) by Cohort – Mathematics

Mathematics: 2013 vs. 2014				
2013 Quartiles	2014 Quartiles			
	Top	Q2	Q3	Bottom
Top	44%	25%	13%	18%
Q2	26%	26%	25%	23%
Q3	17%	28%	31%	24%
Bottom	13%	21%	31%	35%

Mathematics: 2014 vs. 2015				
2014 Quartiles	2015 Quartiles			
	Top	Q2	Q3	Bottom
Top	35%	27%	20%	19%
Q2	27%	28%	26%	20%
Q3	21%	28%	29%	23%
Bottom	18%	18%	25%	38%

Mathematics: 2013 vs. 2015				
2013 Quartiles	2015 Quartiles			
	Top	Q2	Q3	Bottom
Top	34%	23%	21%	22%
Q2	29%	23%	26%	22%
Q3	19%	31%	27%	23%
Bottom	18%	23%	27%	32%

Note: Estimates represent school random effects from Model 7. Our sample includes only fifth grade students who took INVALSI in 2013–2015, for whom we observe second grade scores. We exclude students with missing values on any of the other variables we use as covariates. Our sample is further restricted to schools with at least 25 students satisfying the above criteria and to the same pool of schools across cohorts (2,247 schools).

Table 16. Correlations with SES across Cohorts

Cohort	Mathematics				Italian			
	Model 1		Model 7		Model 1		Model 7	
2013	0.306	*	0.278	*	0.362	*	0.286	*
2014	0.265	*	0.247	*	0.386	*	0.330	*
2015	0.208	*	0.121	*	0.408	*	0.322	*
Differences								
2013 vs. 2014	0.041		0.031		-0.024		-0.044	
2014 vs. 2015	0.057	+	0.126	*	-0.021		0.008	
2013 vs. 2015	0.098	*	0.157	*	-0.046		-0.036	

Note: The top panel displays coefficients from linear regressions of average standardized SES on school performance estimates for mathematics and Italian by cohort using Model 1 and Model 7. All regressions have 2,247 observations with weights equal to each school's number of 5th graders. The bottom panel shows the difference between each correlation with SES across cohorts. We exclude students' SES and school indexes from Model 7, as this model include SES controls. Stars indicate statistical significance at 1 percent and (+) at 5 percent.

Table 17. Descriptive Statistics of school-effects' estimates across Metrics-Italian

Model		Mean	1st Decile	1st Quartile	3rd Quartile	9th Decile
M1	INV	0.032	-0.961	-0.479	0.541	1.056
	WLE	0.029	-0.972	-0.484	0.535	1.056
	TEACH	0.026	-1.075	-0.531	0.587	1.123
M3	INV	0.063	-0.830	-0.345	0.476	0.981
	WLE	0.061	-0.822	-0.354	0.472	0.963
	TEACH	0.043	-1.065	-0.505	0.594	1.109
M7	INV	0.059	-0.947	-0.446	0.540	1.095
	WLE	0.056	-0.942	-0.452	0.534	1.085
	TEACH	0.038	-1.125	-0.559	0.630	1.178

Note: Our sample includes only fifth grade students who took Math INVALSI standardized tests in 2013–2015, for whom we observe second grade scores and teacher assessment scores (*TEACH*). We exclude students with missing values on any of the other variables we use as covariates. Lastly, we restrict our analysis to schools with at least 25 students satisfying the above criteria. Our sample is further restricted to schools with at least 25 students satisfying the above criteria (5,313 schools). We calculate each student's percentile relative to all other students in our sample and cohort to assess stability across metrics.

Table 18. Descriptive Statistics of school-effects' estimates across Metrics-Mathematics

Model		Mean	1st Decile	1st Quartile	3rd Quartile	9th Decile
M1	INV	0.018	-0.985	-0.503	0.516	1.046
	WLE	0.011	-0.994	-0.521	0.512	1.045
	TEACH	0.034	-1.074	-0.520	0.571	1.086
M3	INV	0.067	-0.903	-0.363	0.536	0.961
	WLE	0.060	-0.899	-0.379	0.520	0.966
	TEACH	0.070	-1.041	-0.462	0.631	1.119
M7	INV	0.052	-0.983	-0.452	0.563	1.077
	WLE	0.044	-0.987	-0.464	0.539	1.087
	TEACH	0.052	-1.100	-0.530	0.646	1.202

Note: Our sample includes only fifth grade students who took Math INVALSI standardized tests in 2013–2015, for whom we observe second grade scores and teacher assessment scores (*TEACH*). We exclude students with missing values on any of the other variables we use as covariates. Lastly, we restrict our analysis to schools with at least 25 students satisfying the above criteria. Our sample is further restricted to schools with at least 25 students satisfying the above criteria (5,313 schools). We calculate each student's percentile relative to all other students in our sample and cohort to assess stability across metrics.

Table 19. Pearson Correlation between school-effects' estimates across Metrics by Models

Method	Mathematics			Italian		
	Corr(INV, WLE)	Corr(INV TEACH)	Corr(WLE TEACH)	Corr(INV WLE)	Corr(INV TEACH)	Corr(WLE TEACH)
M1	0.998 *	0.196 *	0.187	0.999 *	0.244 *	0.245 *
M2	0.997 *	0.182 *	0.174 *	0.998 *	0.169 *	0.170 *
M3	0.997 *	0.116 *	0.100 *	0.998 *	0.037 *	0.038 *
M4	0.997 *	0.030	0.022	0.998 *	-0.007	-0.007
M5	0.997 *	0.146 *	0.128 *	0.998 *	0.057 *	0.058 *
M6	0.997 *	0.044 *	0.039 *	0.998 *	-0.004	-0.004
M7	0.997 *	0.033	0.026	0.998 *	-0.011	-0.011

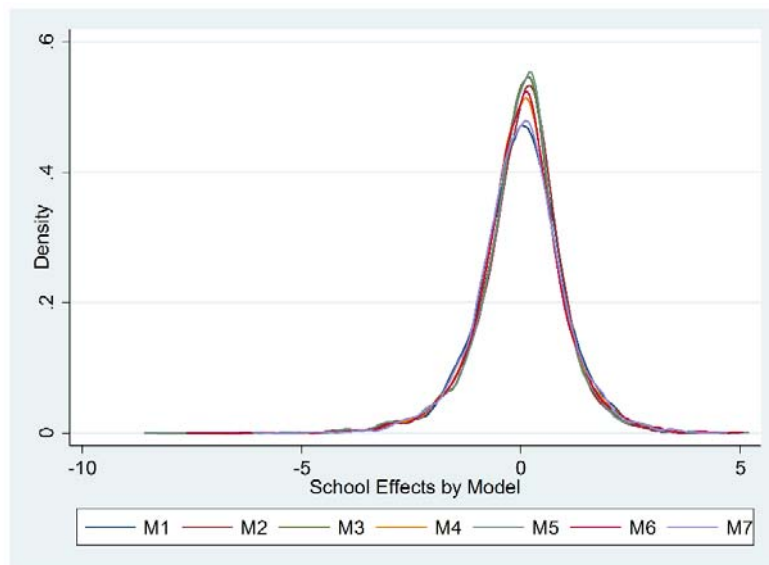
Note: This table displays Pearson correlations across metrics by Model. Our sample includes only fifth grade students who took Math INVALSI tests in 2013–2015, for whom we observe second grade scores and teacher assessment scores (*TEACH*). We exclude students with missing values on any of the other variables we use as covariates. Lastly, we restrict our analysis to schools with at least 25 students satisfying the above criteria. Our sample is further restricted to schools with at least 25 students satisfying the above criteria (5,313 schools). We calculate each student's percentile relative to all other students in our sample and cohort to assess stability across metrics. These correlations use weights equal to each school's number of students. Stars indicate statistical significance at 1 percent.

Table 20. Pearson Correlation between school-effects' estimates across Cohorts-Model 1 and 7

	MODEL 1			MODEL 7		
	INV	WLE	TEACH	INV	WLE	TEACH
Math						
Corr(2013, 2014)	0.416 *	0.418 *	0.569 *	0.303 *	0.286 *	0.520 *
Corr(2013, 2015)	0.317 *	0.334 *	0.571 *	0.146 *	0.157 *	0.510 *
Corr(2014, 2015)	0.361 *	0.361 *	0.553 *	0.223 *	0.223 *	0.486 *
Italian						
Corr(2013, 2014)	0.411 *	0.412 *	0.574 *	0.232 *	0.228 *	0.517 *
Corr(2013, 2015)	0.397 *	0.400 *	0.552 *	0.153 *	0.154 *	0.488 *
Corr(2014, 2015)	0.420 *	0.420 *	0.575 *	0.270 *	0.270 *	0.511 *

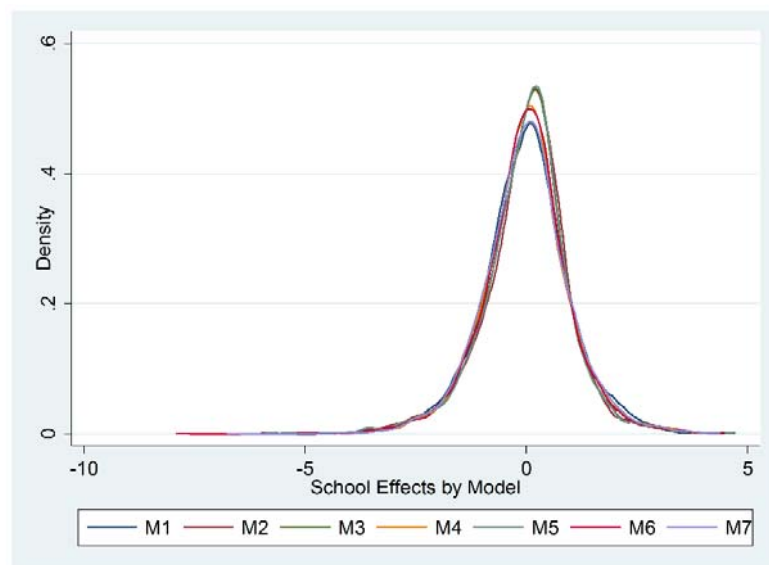
Note: This table displays Pearson correlations across metrics and cohorts for Model 7. Our sample includes only fifth grade students who took Math INVALSI tests in 2013–2015, for whom we observe second grade scores and teacher assessment scores (*TEACH*). We exclude students with missing values on any of the other variables we use as covariates. Lastly, we restrict our analysis to schools with at least 25 students satisfying the above criteria. Our sample is further restricted to schools with at least 25 students satisfying the above criteria (5,313 schools). We calculate each student's percentile relative to all other students in our sample and cohort to assess stability across metrics. These correlations use weights equal to each school's number of students. Stars indicate statistical significance at 1 percent.

Figure 1. Distribution of School Effects' estimates Across Models -Italian

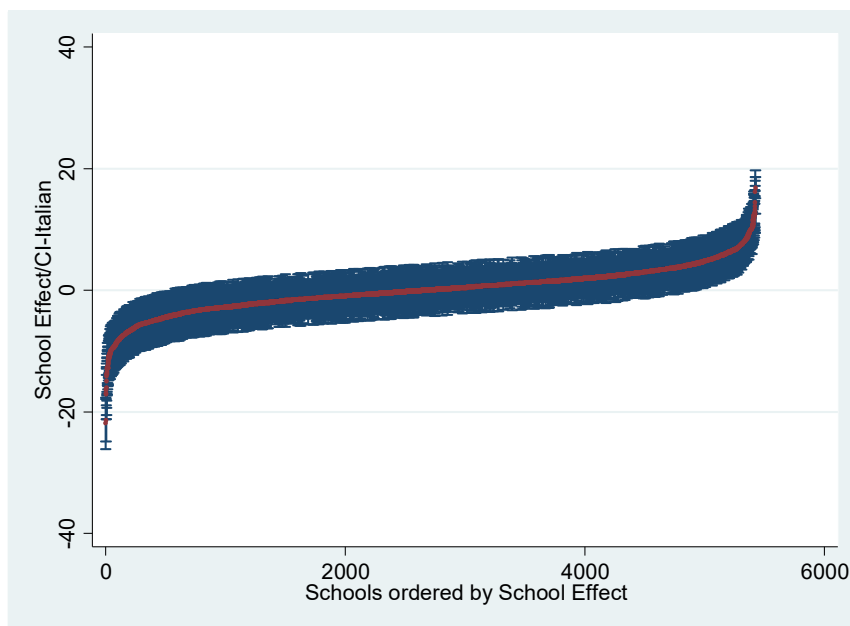


Note: Distribution of school effect estimates across models.

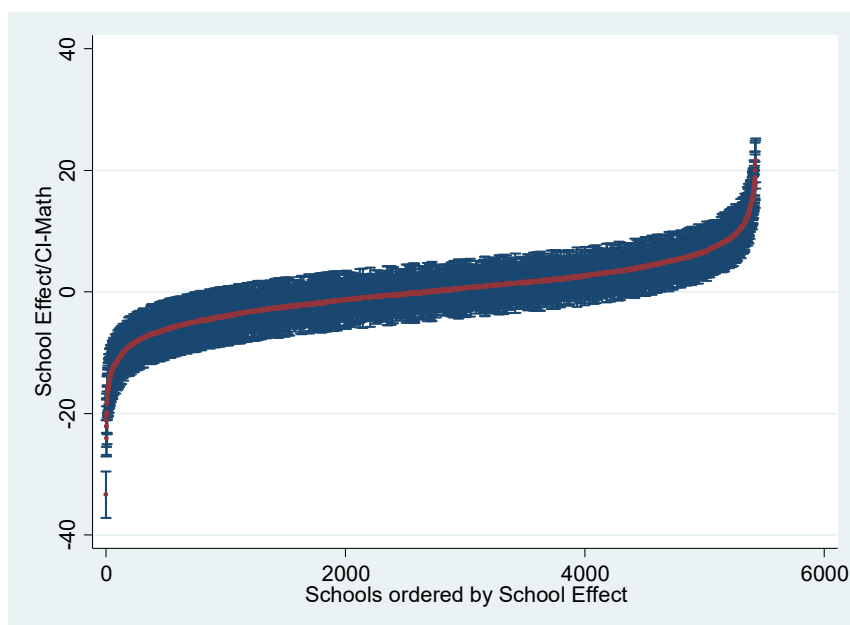
Figure 2. Distribution of School Effects' estimates Across Models -Mathematics



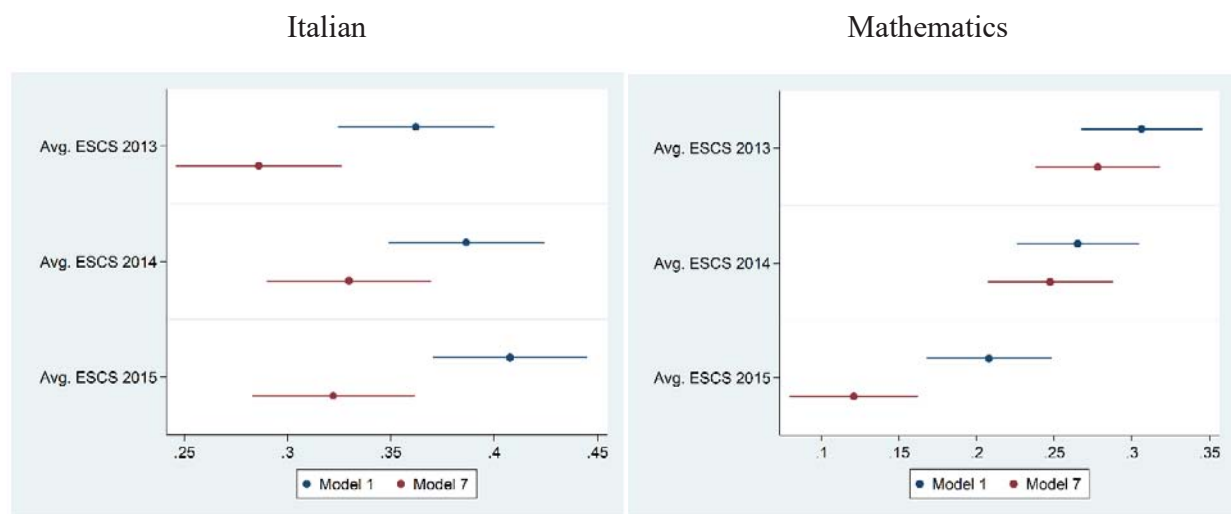
Note: Distribution of school effect estimates across models.

Figure 3. School Effects with Confidence Intervals-Italian

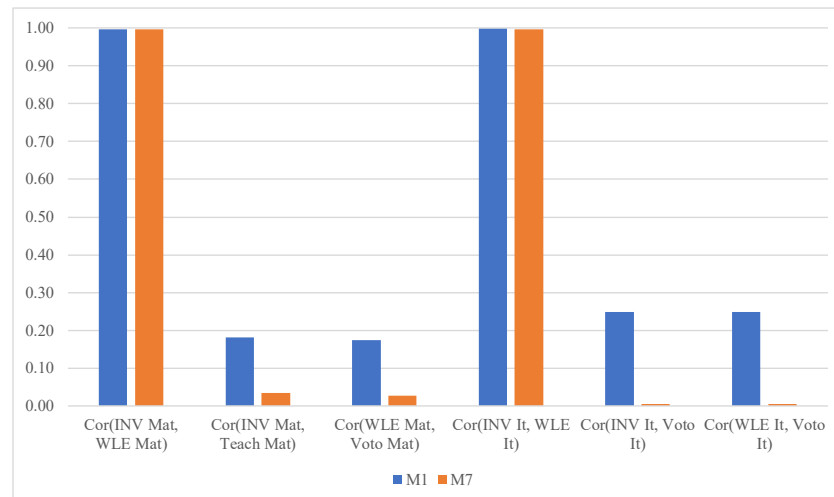
Note: School performance estimates for all cohorts presented with 95% confidence intervals. To avoid standardizing standard errors to construct confidence intervals, we are presenting absolute school effects instead of standardized ones.

Figure 4. School Effects with Confidence Intervals-Mathematics

Note: School performance estimates for all cohorts presented with 95% confidence intervals. To avoid standardizing standard errors to construct confidence intervals, we are presenting absolute school effects instead of standardized ones.

Figure 5. Correlations between school effects, re-estimated with SES, by Model

Note: Correlations between re-estimated school performance estimates and average SES across cohorts. These correlations are statistically significant at 1 percent and use weights equal to each school's number of students.

Figure 6. Spearman Correlations between school-effects' estimates Across Metrics -Model 1 and 7

Note: Spearman correlations between outcome metrics are presented using unadjusted Model 1 and adjusted Model 7.

ANNEX

A Table 1. Differences in Means

Variables	All (S0)		Sample 1 (S1)		Diff. in Means (S0-S1)	Sample 2 (S2)		Diff. in Means (S1-S2)
	Mean	SD	Mean	SD		Mean	SD	
<i>Demographics and SES</i>								
Female	0.49	0.50	0.50	0.50	-0.01	0.50	0.50	0.00
Italian	0.96	0.21	0.97	0.17	-0.01	0.97	0.17	0.00
Regular Student	0.96	0.19	0.98	0.14	-0.02	0.98	0.13	0.00
Average Student SES	0.04	0.96	0.11	0.95	-0.07	0.12	0.94	-0.01
Father in High Income Occupation	0.22	0.42	0.22	0.42	0.00	0.22	0.41	0.00
Mother in High Income Occupation	0.11	0.32	0.11	0.32	0.00	0.11	0.32	0.00
Parents Attended College	0.68	0.47	0.69	0.46	-0.01	0.69	0.46	0.00
<i>Outcome Metrics</i>								
INVALSI Score in Ma, 5th Grade	58.86	18.74	59.71	18.41	-0.85	60.21	18.39	-0.50
INVALSI Score in Ma, 2nd Grade	59.41	20.33	59.86	19.98	-0.45	60.35	19.55	-0.49
INVALSI Score in It, 5th Grade	61.73	17.69	63.38	17.10	-1.65	63.44	16.84	-0.06
INVALSI Score in It, 2nd Grade	65.53	18.79	66.16	18.29	-0.63	67.27	17.98	-1.11
WLE in Math, 5th Grade	202.65	41.14	205.79	39.84	-3.14	206.58	39.12	-0.79
WLE in Italian, 5th Grade	200.60	41.52	203.88	39.69	-3.28	205.25	38.54	-1.37
Teacher's Score in Math, 5th Grade	7.84	1.14	7.94	1.10	-0.10	7.95	1.09	-0.01
Teacher's Score in Italian, 5th Grade	7.78	1.09	7.87	1.04	-0.09	7.87	1.03	0.00
Number of Students	1,426,539		790,519			482,379		
Number of Schools	7,242		5,427			2,247		

Note: Average INVALSI and WLE scores are corrected by the propensity to test score manipulation using INVALSI estimations. Students' SES is measured by the ESCS. The ESCS is an index of economic, social and cultural status that was provided by INVALSI and was estimated using availability of household assets, parental education and employment status. We used oral scores for teacher assessments because written examinations had 80 percent of missing values. The first two columns report averages and standard deviations for the population of students who were tested by INVALSI from cohorts 2013 to 2015 (S0). Sample 1 (S1) includes only fifth grade students who took INVALSI in 2013–2015, for whom we observe second grade scores. We exclude students with missing values on any of the other variables we use as covariates. Lastly, we restrict our analysis to schools with at least 25 students satisfying the above criteria. Sample 2 (S2) further restricted to schools satisfying S1 criteria and to the same pool of schools across cohorts.

ATable 2. Modeling Decisions – covariates included in the various models

Method and Covariate Choice							
	Student's Second Grade Score	Student Demographic Control	Class Means	Class Fixed Effects	School Means	School Fixed Effects	School Random Effects
Model 1							
Model 2	X					X	
Model 3	X	X				X	
Model 4	X	X	X			X	
Model 5	X	X		X			
Model 6	X	X	X		X		
Model 7	X	X	X				X

Note: Models 2-7 control for controls for demographic characteristics such as dummies for student type (i.e. regular, overage, and underage), gender, immigration status, and student SES measured by ESCS student index. All models account for test score manipulation using INVALSI's propensity score to cheating, and cohort dummies. Classroom means include classroom-level means of student gender, immigration status, student type, as well as the class size and average second grade scores.

Table 3. Transition Matrices of School Quartiles (based on school-effects' estimates) for Italian (top panel) and Mathematics (bottom panel)

Italian: Model 7 vs. Model 3				
M3 Quartiles	M7 Quartiles			
	Top	Q2	Q3	Bottom
Top	89%	10%	0%	0%
Q2	11%	76%	13%	0%
Q3	0%	14%	76%	11%
Bottom	0%	0%	11%	89%

Mathematics: Model 7 vs. Model 3				
M3 Quartiles	M7 Quartiles			
	Top	Q2	Q3	Bottom
Top	87%	12%	1%	0%
Q2	13%	72%	14%	1%
Q3	0%	16%	73%	11%
Bottom	0%	0%	12%	88%

Note: Adjusted school effects' estimates are presented by quartiles from from Model 3 and school random effects from Model 7. Our sample includes only fifth grade students who took INVALSI in 2013–2015, for whom we observe second grade scores. We exclude students with missing values on any of the other variables we use as covariates. Our sample is further restricted to schools with at least 25 students satisfying the above criteria (5,427 schools).

ATable 4. Comparison between INVALSI Value Added Estimates and Our Sample (based on school-effects' estimates) for Italian (top panel) and Mathematics (bottom panel)

Italian		Positive	Average	Negative	Total
Positive	N	285	157	19	461
	%	61.82	34.06	4.12	100
Average	N	304	3361	310	3,975
	%	7.65	84.55	7.8	100
Negative	N	67	125	300	492
	%	13.62	25.41	60.98	100
Total		656	3,643	629	4,928
		13.31	73.92	12.76	100

Mathematics		Positive	Average	Negative	Total
Positive	N	423	116	6	545
	%	77.61	21.28	1.1	100
Average	N	302	3309	317	3,928
	%	7.69	84.24	8.07	100
Negative	N	10	102	343	455
	%	2.2	22.42	75.38	100
Total		735	3,527	666	4,928
		14.91	71.57	13.51	100

Note: 4,075 (83 percent) and 3,946 (80 percent) schools out of 4,928 are estimated to be on the same category (i.e. positive, average or negative value added) according to INVALSI and our estimation using our sample restriction for both Italian and Mathematics, respectively. Our estimation includes only fifth grade students who took INVALSI in 2015, for whom we observe second grade scores. We exclude students with missing values on any of the other variables we use as covariates. Lastly, we restrict our analysis to schools with at least 25 students satisfying the above criteria.

Table 5. Correlations between Linear and Non-Linear Specifications for Mathematics and Italian

Model	Mathematics	Italian
Model 2	0.999 *	0.999 *
Model 3	0.998 *	0.999 *
Model 4	0.999 *	0.999 *
Model 5	0.999 *	0.999 *
Model 6	0.999 *	0.999 *
Model 7	0.999 *	0.999 *

Note: We fit the same models using both a linear and a fifth-order polynomial of relationship between fifth grade and second grade scores. These correlations use weights equal to each school's number of students. Stars indicate statistical significance at 1 percent.