PROGRAMMA EDUCATION
FGA WORKING PAPER
N. 47 (04/2013)

Patterns of Value-Added Creation in the Transition from Primary to Lower Secondary Education in Italy

Gianfranco De Simone e Andrea Gavosto
Fondazione Giovanni Agnelli

Φ
Fondazione
Giovanni Agnelli

www.fondazione-agnelli.it **segreteria@fga.it**

PATTERNS OF VALUE-ADDED CREATION IN THE TRANSITION FROM PRIMARY TO LOWER SECONDARY EDUCATION IN ITALY

Gianfranco De Simone - *Fondazione Giovanni Agnelli*

Andrea Gavosto - *Fondazione Giovanni Agnelli*

April 2013

Abstract

We estimate a model of cognitive gain by exploiting a unique dataset that tracks the performance over time in Reading and Math of over 6000 students in three Italian provinces. The data were gathered within the first experiment conducted in Italy aiming at assessing the value added provided by schools on the basis of longitudinal information on students. Among the 72 schools involved in the experiment, we are able to identify those which, on average, add more value to the achievements of their students during the crucial transition from primary to lower secondary (from grade 5 to grade 6). We also explore how best performing schools make a difference by narrowing achievement gaps usually associated with individual characteristics of students (gender, socio-economic background, foreign origin). On a more general basis, we are able to show that a considerable share of variability in value-added creation lies at the level of classes within school. Although reduced in scope, such class-level difference in value-added creation persists once we control for class composition in terms of observable student characteristics (level and heterogeneity of socio-cultural background). The remaining part of the unexplained variability provides an estimate of the joint effect of teacher effectiveness and other class-level unobserved factors.

Key-words: Value added, Lower-Secondary Education, Heterogeneity of Effects, Class Formation, Teacher Effectiveness

JEL classification: C23, I2

1. Introduction

Since the 1966 Coleman report, student scores at standardized tests have been increasingly used in order to assess the effectiveness of a school or of an individual teacher, but only recently they have become an essential tool to implement accountability schemes into education systems. Following the experience of countries such as England over the 1990s, the No Child Left Behind Act of 2001 has imposed to states in US to test students annually in grades 3-8 and in one grade in high school.

The availability of such a wealth of data on achievements has helped to develop new models and techniques which attempts to address school accountability, instructional improvement and parents' choice. Initially, models relied on the use of raw achievement data[1]; however, it became immediately clear that school outcomes were largely influenced by family socio-economic conditions (McCall, Kinsbury and Olson, 2004). This led to the development of contextualized attainment models, based upon large cross-sections of data, which included measures of the socio-economic context (Aitkin and Longford, 1986; Goldstein, 1986; Willms and Raudenbush, 1989). Albeit an improvement, contextualized attainment models lacked information on students' ability and prior achievements, which can explain a good deal of individual performances. Hence, value-added models, which tracks individual test scores over time, became increasingly popular in England (FitzGibbon, 1997) and in the US (Sanders, Saxton and Horn, 1997). Education is a cumulative process, though: therefore, the context can affect both the level and the rate of growth of each individual learning (Ballou et al., 2004). For this reason, in this paper we will use variants of a contextualized value-added model, which includes both prior achievements and student background characteristics as predictors of current performance.

Many researchers have questioned the validity of the inferences drawn from value-added models in view of the many technical challenges that exist: accuracy of the data, linkage of tests carried out at different grades, bias in estimates and measurement errors (see Schmidt et al., 2005,Rothstein, 2009, Reckase, 2008). Also, it is well known that value-added can induce distorted incentives for teachers and principals, such as teaching to the test (Khon, 2000; Nichols and Berliner, 2005). Most of the argument against using value-added for evaluation purposes arises because, in states such as California, tests have been used to assess the contribution of individual teachers on the basis of their students' performance over the years. Opponents argue that this exercise lacks sufficient precision (Rothstein,

[1] See Oecd (2008) for a survey of applications.

2010) and, as a consequence, the policy to shame teachers who are reportedly ineffective can lead to gross misjudgment.

In the school year 2010/11, the Italian Ministry of Education has launched a pilot program aiming at assessing the effectiveness of lower secondary schools in three provinces (Pavia, Arezzo and Siracusa) on the basis of the cognitive progress of their students. In such scheme, the value added is intended to be computed at the level of the school unit: the underlying idea is that individual contribution cannot be disentangled and what matters is the result of a team work (Bertola and Checchi, 2008). Notwithstanding this cautious choice, this first attempt to employ standardized test scores for the evaluation of schools' performance has been initially greeted with a lot of concern by Italian teachers.

In this paper we try to get as much information as possible out of value added assessment of the 72 schools (six thousands pupils) involved in the program. Our aim is to offer a wider view of what can be learned on educational quality from such a measure of school performance. We use longitudinal data on individual performances in reading and math from standardized tests and we estimate measures of cognitive gain. Once a number of contextual and individual factors are controlled for, the variation and the of cognitive progress are analyzed along several dimensions.

The contribution of this paper is twofold. On the one hand, for the first in Italy, we are able to explore different specifications and econometric techniques in order to define as precisely as possible the value-added created by schools in our sample, starting from standardized test scores available within the National Assessment System set up by the National Institute for the Evaluation of the Education System (INVALSI). On the other hand, we take a first stab at what are the main features of the best- and worst-performing schools. Furthermore, by means of a variance decomposition technique, we are able to identify the share of variability in achievements attributable to the quality of teaching and management and to other class- and school-level factors.

The paper is organized as follows. In the section 2, we will explain the main features of the evaluation experiment within which the data were collected. Section 3 provides different estimates of the cognitive gain function for the schools in the sample. In section 4, we will look at separate regressions for best- and worst- performing schools and decompose the overall variance in cognitive gains in order to infer some hints of what are the main characteristics of the best schools. Conclusions follow in section 5.

2. The case study: 72 schools in three Italian provinces

Italy is the only developed country which lacks a system of evaluation of schools' and teachers' performances. Moreover, only recently did the country adopt standardized tests to monitor the cognitive achievement of students. The assessment is administered by INVALSI, the agency of the Ministry of Education which runs compulsory reading and mathematics tests for all the students' population in grades 2, 5, 6, 8 and 10.

In 2010 the then Ministry of Education decided to run two distinct pilot programs for the assessments of schools and teachers. The projects, which stirred a hot public debate, were aimed at experimenting two different models of evaluation to be applied later to all schools and teachers in the country.

In this paper we focus on the pilot assessment of school performances. According to the plan devised by a group of experts, the evaluation scheme spans over three years and applies to the Italian lower secondary schools (grade 6-8, corresponding to students of 11 to 14 years of age). The project was launched in three Italian provinces - one in the North (Pavia), one in the Centre (Arezzo) and one in the South (Siracusa) of the country - in order to achieve some regional balance. All schools in the selected provinces were contacted by the Ministry but only 72 out of 123 accepted to join the experiment (20 in Pavia, 14 in Arezzo, 38 in Siracusa). Five additional schools from a fourth northern province (Mantua) managed to join the group, but are not included in our sample. Although it is reasonable to expect that some sort of self-selection occurred, a probit estimate computed by Checchi-De Simone-Rettore (2013) on the same sample reveals that the two groups of school (participating, not participating) are balanced across several observable characteristics[2]. Thus, self-selection does not appear to be a relevant issue in our sample.

The evaluation scheme relies on two main pillars. One is a measure of contextual value added, based upon the results for each individual student of the INVALSI tests in grade 5 (entry point) and grade 6 (end of the first year of lower secondary school); eventually the same students will be followed up to the 8th grade (end of lower secondary school). This is the measure we will focus on in this paper. The other pillar of the experiment consists of on-site visits by teams of three external experts led by a high ranking official of the Ministry. The objective is to assess the quality of school performance in domains not necessarily captured by standardized test of students’ achievement: practices of inclusion

[2] See Table A1 in the Appendix.

of immigrant and disabled students, support of academically weak pupils and enhancement of academically excellent ones, support of students in the last year for the choice of the high school, innovative practices of self-assessment and pupils' evaluation. Inspectors have a checklist of good practices that schools are expected to have undertaken in each of these seven domains: if a school fulfills all of them, it is graded at the top (4 out of 4) in that particular domain; if it has not undertaken any of them, the grade is zero. Grades in the seven domains are averaged (with a weight of 40% overall) together with the value-added scores in Italian (35%) and in Math (25%) and schools are ranked within each province.

The top 25% schools in each province received a monetary award which amounts to 35,000 euros. This is just the first installment of the overall prize (100.000 euros) which will be delivered at the end of the third year of the lower secondary cycle (8th grade), when all schools which joined the experiment will be tested again, on the basis of both value-added and on-site visits. In the meantime, after the 6th grade test, all the schools in the sample have received a detailed report which describes their strengths and weaknesses, so that they can start a training programme for teachers.

The purpose of the experiment is twofold: on the one hand, it attempts to create a fully-fledged system of school evaluation, based upon measures of value added; on the other hand, it purports to see how do schools react to monetary incentives and whether these elicit a greater effort by teachers and principals[3]. In this paper we will make use of the first leg of the experiment, the one conducted in 2011 between 5th and 6th grade, to investigate what kind of value added measures can be estimated with available data and what patterns of value-added formation prevail in our sample.

2.1. The data

Descriptive statistics of the dataset are reported in Table 1. We have two scores in reading and math from INVALSI tests over two points in time for nearly 9 students out of 10 of those attending schools involved in the project (88.2% in reading, 89.5% in math). It took some work to recover the prior scores (INVALSI test scores at 5th grade): in fact, due to a bizarre interpretation of the Italian privacy

[3] In particular the exercise aims at detecting 3 effects of the monetary incentives to schools: 1) whether schools in provinces where evaluation has been carried out perform better than schools in other areas; 2) whether schools that have received the initial prize carry their effort along until the third year in order to preserve their place in the ranking or they rather lose track vis-à-vis schools that have initially come close to the top ; and 3) since schools are free to share the award among teachers in the way they prefer, whether schools where prizes have been shared in a more equitable way among teachers perform better. See Checchi-De Simone-Rettore (2013) for some early quasi-experimental evidence on the short term impact of the program.

law by the relevant authority, neither the Ministry of Education nor INVALSI itself were allowed to identify the name of the students who carried out the test, but could only keep track of his/her digital code. Hence, schools had to match the name and the code of respective students and thus create a longitudinal database. Fortunately, most of the participating schools are vertically integrated (offering both primary and lower secondary education) with a common administrative office that could provide and link achievements of individuals over time. Still, a number of records could not be matched, which explains why value-added has been computed for less than the 100% of students[4] involved.

Table 1: Descriptive statistics

Level	Variable	Obs	Mean	Std. Dev.	Min	Max
Student	Test score at grade 6 – Reading	5987	0.00	1.00	-3.56	2.07
	Test score at grade 6 – Math	5833	0.00	1.00	-2.46	3.04
	Test score at grade 5 – Reading	5284	0.00	1.00	-3.22	1.72
	Test score at grade 5 – Math	5220	0.00	1.00	-3.00	2.02
	Female	6018	0.49	0.50	0.00	1.00
	ESCS	6015	-0.05	0.98	-3.31	2.45
	Grade repeater in primary school	6018	0.08	0.26	0.00	1.00
	1st generation immigrant student	6003	0.07	0.25	0.00	1.00
	2nd generation immigrant student	6003	0.04	0.20	0.00	1.00
School	Province	6019				
	Arezzo	899	0.15		0.00	1.00
	Pavia	2557	0.42		0.00	1.00
	Siracusa	2563	0.43		0.00	1.00
	Small town	6019	0.80	0.40	0.00	1.00
	Not vertically integrated with a primary school	6019	0.26	0.44	0.00	1.00
	Number of school units to be managed	6019	3.64	2.31	1.00	11.00
	Share of teachers with temporary contract	6019	0.20	0.15	0.00	0.85
	Involved in program PQM - Reading	6019	0.06	0.17	0.00	1.00
	Involved in program PQM - Math	6019	0.06	0.16	0.00	1.00
	Involved in program Mathabel	6019	0.03	0.13	0.00	1.00
	Average test score at grade 5 - Reading	6019	-0.07	0.40	-2.24	0.61
	Average test score at grade 5 - Math	6019	-0.10	0.40	-2.22	0.76
	Average ESCS	6019	-0.07	0.30	-1.00	0.66
	Share of disabled students	6019	0.04	0.03	0.00	0.12
	Share of immigrant students	6019	0.11	0.08	0.00	0.28

[4] Missing prior scores are not necessarily existing scores that have not been matched. In fact, the 5th grade test is not used for grading students and thus it is not replicated for students missing on the day in which it is administered. On average, the share of students missing the test is around 3-4% of the relevant cohort.

We also have information about the gender, the socio-cultural background, the nationality and the regularity in the course of study of students. At the school level we have information related to: the location of schools (province, big city vs. small town), the organizational complexity of the school (number of separate units to be managed, possible vertical integration with a primary school), the share of teachers with temporary contracts. We also know if the schools have been involved in supporting programs by either the Ministry of Education or the European Union (PQM, Mathabel).

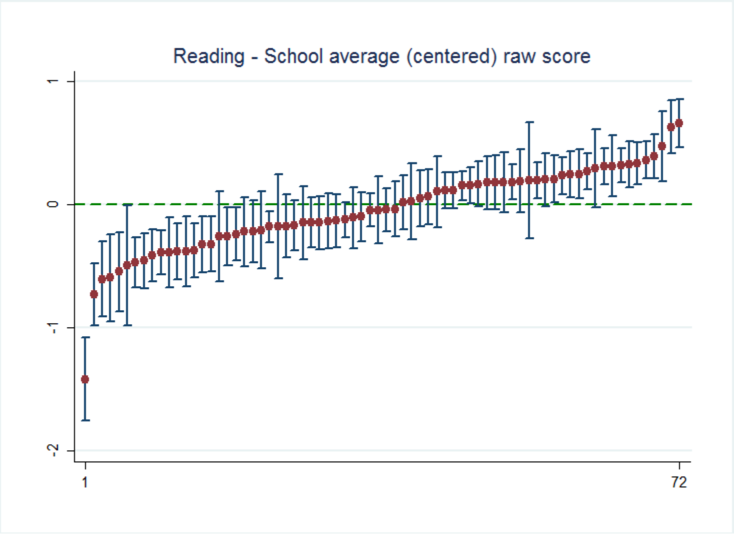
2.2. Dealing with anomalous observations

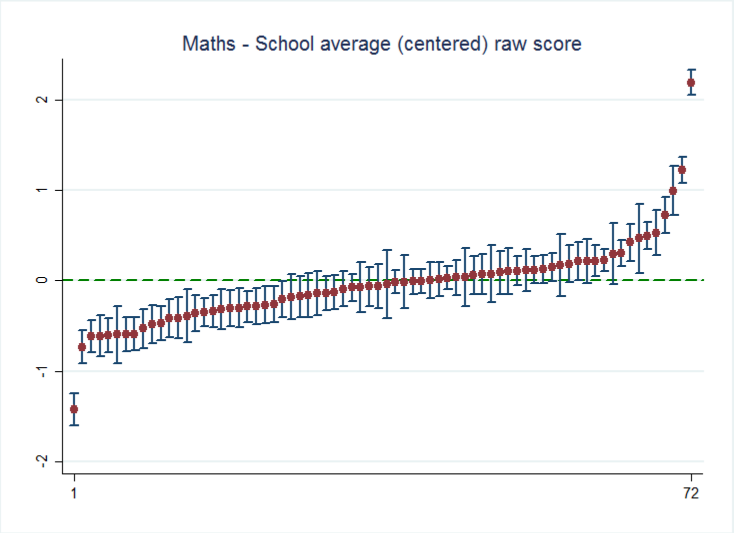
The inspection of school average raw scores in reading and math reveals the presence of a few odd observations (Figure 1). As the range of performances spans from -1 to +1, we observe two schools that lie significantly above the upper limit in math and a single school that reports a score below the lower limit in both math and reading. A fourth school reports a math score on the upper limit (.99) with a large confidence interval. It is hard to identify the origin of such extreme cases as they may depend on errors in the data collection as well as on opportunistic behavior (cheating) in some school[5].As we are dealing with a small sample and a single figure outside the range can affect estimates substantially, we decided to drop the outliers to ensure that our results are not driven by extreme values. More specifically, we leave out the worst performing school (extreme left of the distribution) in both reading and math and the three top performing schools in math[6] (extreme right).

[5] To prevent schools from adopting opportunistic behaviors, external examiners were sent to check tests administration in each class involved in the project. As a consequence the risk of cheating should be limited.

[6] We adopt a conservative strategy and we drop also the school on the upper limit. In fact, for that school, we could have observed a score over the upper limit (1) with a probability of 50% .

Figure 1: Distribution of school raw scores at grade 6 in reading and math





2.3. Cognitive gain when scale is missing

INVALSI test scores at different grades are not vertically linked by a common scale. This makes grade to grade progress difficult to measure (Young, 2006). Plain value-added as the difference between test scores is therefore impossible to compute for Italian students and schools[7]. As an alternative strategy, scholars tend to adopt models of cognitive gain where previous scores of students

[7] Martineau (2006) shows that vertical scaling can be itself a source of bias in the estimates of value added. Tong and Kolen (2007) discuss how achievement measures may vary as the methodology applied for scaling varies.

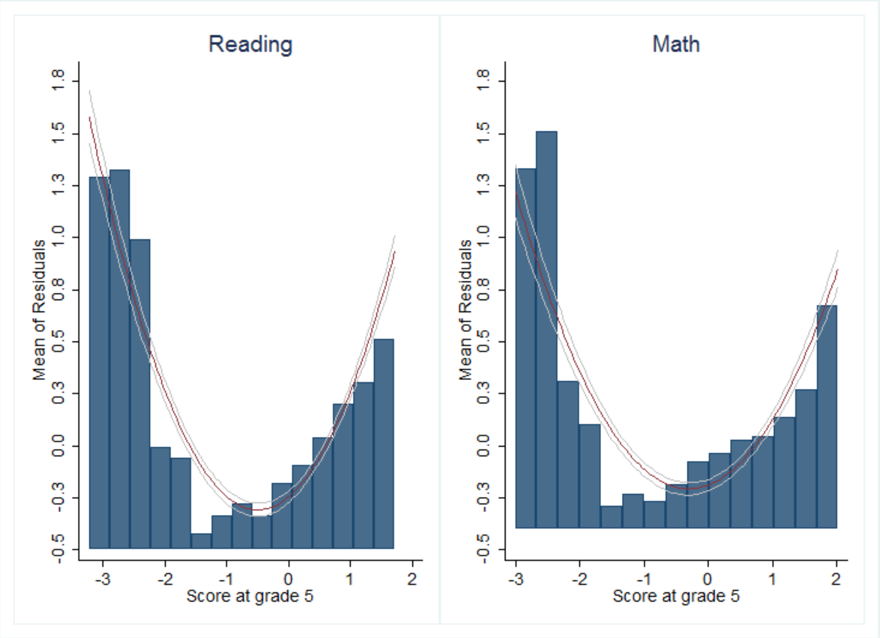
are used as predictors of current achievements. However, two subsequent scores are not necessarily linked through a first-order linear relationship such as the following:

$$s_{i,j}^{m,6} = \alpha_0 + \alpha_1 s_{i,j}^{m,5} + r_{i,j}^{m}, \qquad (1)$$

where $s_{i,j}^{m}$ represents the achievement of a student *i* of a school *j* in subject *m* (in our case, reading and math) over two points in time (in our case, grade 5 and 6) and $r_{i,j}^{m}$ is a residual term.

By looking at the residuals of two separate estimates of equation (1) for reading and math on our sample of students it appears that INVALSI test scores at grade 5 and 6 are linked through a non-linear relationship (Figure 2). The U-shaped distribution of cognitive progress (residuals) across the entry levels of students (scores at grade 5) suggest that the proper functional form linking the scores at the two grades should be polynomial of order 2.

Figure 2: Distribution of cognitive progress as defined in equation (1) by scores at grade 5

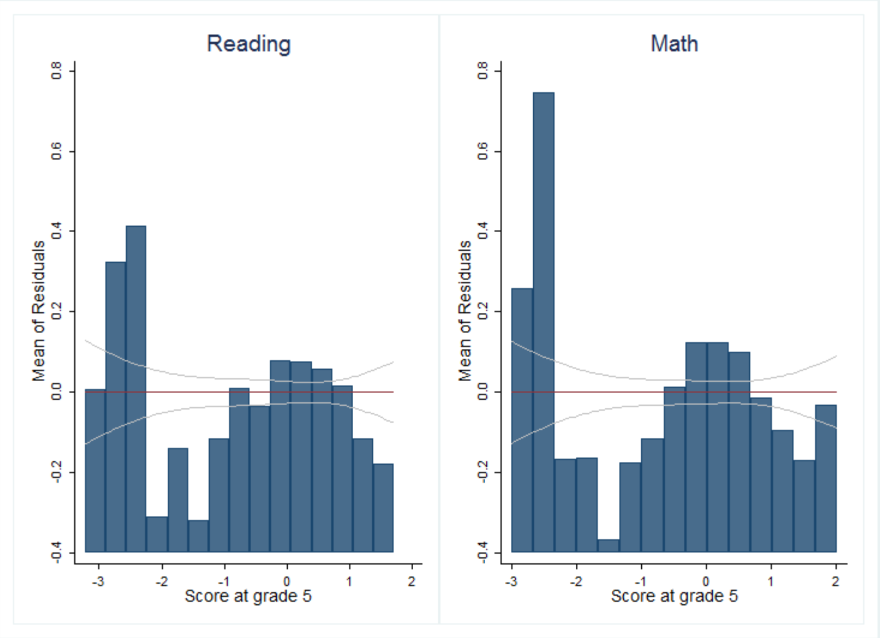


So an unadjusted model of cognitive progress able to capture the link between INVALSI test scores at grade 5 and 6 would take the following functional form:

$$s_{i,j}^{m,6} = \alpha_0 + \alpha_1 s_{i,j}^{m,5} + \alpha_2 \left(s_{i,j}^{m,5}\right)^2 + r_{i,j}^{m}. \qquad (2)$$

As expected, the data reveal that the distribution of cognitive progress estimated by (2) shows no clear pattern of association with the scores at grade 5 (Figure 3).

Figure 3: Distribution of cognitive progress as defined in equation (2) by scores at grade 5



3. Assessing average school cognitive gain

3.1. The model: simple linear model, school-level fixed effects or multilevel mixed-effects?

Equation (2) does not lead to a fair comparison between schools: students characteristics and other school-level contextual factors which impinge on the achievements of students are in fact exogenous to schools. Thus we need to adjust our estimates of cognitive progress for all observable characteristics of students and external factors that may affect the educational process but are not directly managed by schools.

In a linear model, this is easily done by including the relevant controls in the specification:

$$s_{i,j}^{m,6} = \alpha_0 + \alpha_1 s_{i,j}^{m,5} + \alpha_2 \left(s_{i,j}^{m,5}\right)^2 + X_i' \beta_1 + Z_j' \beta_2 + u_{i,j}^{m}, \qquad (3)$$

where X'_i is a vector of student and family characteristics and Z'_j is a set of contextual factors affecting the activity of schools. Value added at the school level is computed as the average of residuals $(u_{i,j}^{m})$, namely the difference between observed achievements and predicted achievements obtained by fitting equation (3):

$$VA_j = ave_i\left(u_{i,j}^{m}\right) = ave_i\left(s_{i,j}^{m,6} - \hat{s}_{i,j}^{m,6}\right) \qquad (3').$$

Such a linear model has the advantages of simplicity, but in order to yield consistent estimates we need to make sure that included covariates are not correlated with the residual term; furthermore, the hypothesis of i.i.d. in the normal distribution of errors should not be violated. Both assumptions are hard to be upheld by the data.

To relax the former and deal with the possible omitted variable bias on the estimated coefficients in equation (3), a model that includes school-level fixed effects can be used, such as:

$$s_{i,j}^{m,6} = \alpha_0 + \alpha_1 s_{i,j}^{m,5} + \alpha_2\left(s_{i,j}^{m,5}\right)^2 + X_i^{'}\beta_1 + \sum_j \gamma_j + u_{i,j}^{m}, \qquad (4)$$

where γ_j is a set of dummies capturing all observed and unobserved factors operating at the school level. Such a specification leads to consistent estimates of coefficients on other factors $(\alpha_0, \alpha_1, \alpha_2, \beta_1)$ but presents some shortcomings: the value added of schools is now mixed up with other school level characteristics in the fixed effect term, and it is hard to tell it apart from all of those school-level contextual factors that should be controlled for in order to compare school performances. Also, as the number of schools increases, the number of school fixed effects to be estimated increases as well and with small samples the estimated fixed effects become unreliable.

A hierarchical (or multilevel) class of models allows to adopt a random effects approach, which on paper is more efficient, as it has a structure of errors which is apparently closer to the reality of schools. In fact, students are grouped into classes that are in turn nested into schools. Mulilevel models yield more accurate estimates of the variability to be attached to the estimates of school value-added. However, to be consistent, random effect estimates require school effects to be uncorrelated with the explanatory variables at the school level: a condition which is often hard to fulfil.

A typical formulation of such models is:

$$s_{i,j}^{m,6} = \alpha_0^{j} + \alpha_1 s_{i,j}^{m,5} + \alpha_2\left(s_{i,j}^{m,5}\right)^2 + X_i'\beta_1 + Z_j'\beta_2 + u_{i,j}^{m} \qquad (5)$$

where

$$\alpha_0^j = A + \varepsilon_0^j. \tag{5'}$$

Residuals at both levels (i, j) are assumed to be i.i.d. and normally distributed, but in this model each school effect consist of two parts: a grand mean (A) and an idiosyncratic component (ε_0^j) that reveals how the school makes a positive or negative difference from the average on its own students' achievements. Thus the deviations from the general mean (ε_0^j) are taken as estimates of school value-added. Notice that we include school-level contextual factors in the fixed portion of equation (5) to produce a fair comparison between schools. Since the school impacts on student performances are treated as random variables, the estimated value-added 'shrinks' toward the general mean and can be considered as a weighted average of the ordinary least squares estimate within each school and of the estimate between schools. Although biased, shrinkage estimates typically have smaller mean squared error than ordinary least squares estimates.

We estimate the three models (linear model, fixed effects and multilevel mixed-effects) in order to investigate their differences. Among student-level regressors (X_i) we include: gender, socio-cultural background, place of birth and nationality (natives / 1st generation immigrant students / 2nd generation immigrant students), possible grade retention at the primary school. These are usual controls in contextualized value-added models. The choice of school-level controls (Z_j) is somewhat more complicated as we want to include just those factors over which the school has no direct control. Therefore we include information related to the location of schools (big city vs. small town, province) to account for both the larger amount of educational resources available in large urban centers (Young, 1998) and the huge gap between the northern and the southern part of the country (Montanaro, 2008; Ferrer-Esteban, 2010). We also control for the organizational complexity of the school as proxied by the number of separate units to be managed by the principal and a dummy that equals 1 when the lower secondary school is vertically integrated with a primary school. The negative impact of instruction discontinuity on students is captured by the presence of teachers with temporary contracts, which leads to greater turnover (Barbieri et al., 2007). We include a dummy when schools receive extra-resources from supporting programmes by the Ministry of Education or the European Union. We also include a number of student characteristics, averaged within each school, to capture the advantage or disadvantage on each single student achievement deriving from the features of the schoolmates (share of disabled students, share of students with a foreign origin, average socio-cultural status, average performance of student at grade 5).

3.2. Results

We report in Table 2 the results obtained by estimating 4 models for each of the two subjects. We start with the unadjusted linear model (column a) and proceed rightward with the adjusted linear model (column b), the school fixed effect model (column c) and the adjusted multilevel model with school random effect (column d), which represent the empirical counterpart of equations (2), (3), (4), (5, 5’) respectively.

The first thing to be noticed is the stability of the estimated coefficients across different specifications. In particular, coefficients on individual variables in the upper block do not vary as we move from the adjusted linear model (b) to the fixed effect model (c): as a consequence, omitted variables bias is not an issue in our sample. Hence, the assumption $E\left(X_i' u_{i,j}^m\right) = 0$ is not violated and equation (3) can be consistently estimated by simple ordinary least squares (OLS). Coefficients on the student-level variables remain stable also when we adopt a multilevel specification with random effects. The main difference between results in columns (b) and (d) is the significance of school-level variables (Z_j). This is not surprising as a considerable portion of variability of these regressors is captured by random school effects (ε_0^j). As discussed above, the model in column (d) has the advantage of relaxing the assumption on the structure of error maintained in linear models; however, the plausible correlation between covariates and random effects can introduce a bias into the estimation of school effects.

We take the adjusted linear model, which delivers consistent estimates, as our preferred specification and compute the school effects via equation (3’). In Table 3 we group the schools on the basis of their effectiveness in both subjects. If we build a 5% confidence interval around their average value-added, we can single out the best performing schools (those significantly above the sample average), the schools performing on the average (not significantly different from the sample average) and the worst performing schools (significantly below the sample average). Only a half of the schools in our sample present the same degree of effectiveness in reading and math. The remaining schools present a significant divergence in effectiveness and thus in the quality of instruction over the two subjects. This results suggests that there could be a large variation across classes and teachers within schools: we will treat this subject in more depth in Section 4.

Table 2: Models of cognitive gain in reading and math (grade 5 to 6)

	Reading				**Math**			
	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
Dep. Variable: Test score at grade 6	*Unadjusted LM*	*Adjusted LM*	*LM with fixed effects*	*Adjusted MLM with random effects*	*Unadjusted LM*	*Adjusted LM*	*LM with fixed effects*	*Adjusted MLM with random effects*
Test score at grade 5	0.658***	0.606***	0.608***	0.608***	0.591***	0.569***	0.574***	0.573***
	[0.0140]	[0.0151]	[0.0148]	[0.0141]	[0.0137]	[0.0146]	[0.0143]	[0.0132]
Test score at grade 5 – squared	0.254***	0.217***	0.219***	0.219***	0.196***	0.179***	0.187***	0.185***
	[0.00929]	[0.00943]	[0.00957]	[0.00872]	[0.00987]	[0.0102]	[0.0107]	[0.00910]
Female student		0.172***	0.174***	0.174***		-0.0284	-0.0157	-0.0175
		[0.0215]	[0.0209]	[0.0209]		[0.0229]	[0.0225]	[0.0222]
Immigrant status (Ref. Native with Italian parents)								
1st generation immigrant student		-0.168***	-0.173***	-0.171***		-0.0145	-0.0178	-0.0166
		[0.0566]	[0.0552]	[0.0495]		[0.0526]	[0.0508]	[0.0519]
2nd generation immigrant student		-0.0958	-0.106*	-0.105**		-0.0997*	-0.107**	-0.105*
		[0.0615]	[0.0600]	[0.0526]		[0.0533]	[0.0526]	[0.0551]
Grade repeater in primary school		-0.203***	-0.186***	-0.188***		-0.144***	-0.136**	-0.137***
		[0.0640]	[0.0610]	[0.0509]		[0.0549]	[0.0539]	[0.0532]
ESCS		0.166***	0.167***	0.167***		0.148***	0.148***	0.148***
		[0.0123]	[0.0119]	[0.0117]		[0.0129]	[0.0125]	[0.0124]
Share of disabled students		-0.00519		-0.00158		0.0102*		0.0136
		[0.00527]		[0.0132]		[0.00571]		[0.0147]
Share of immigrant students		0.00114		0.00110		-0.00146		4.37e-05
		[0.00258]		[0.00718]		[0.00266]		[0.00759]
Average ESCS		0.225***		0.186		0.144**		0.207
		[0.0522]		[0.125]		[0.0565]		[0.147]
Average test score at grade 5		-0.296***		-0.258***		-0.264***		-0.295***
		[0.0374]		[0.0792]		[0.0407]		[0.0895]
Share of teachers with temporary contract		0.000869		0.000818		0.00239*		0.00432
		[0.00135]		[0.00262]		[0.00125]		[0.00274]
Involved in program PQM		0.00182***		0.00157		0.00236***		0.00305*
		[0.000707]		[0.00167]		[0.000795]		[0.00183]
Involved in program Mathabel						0.000711		0.00120
						[0.000918]		[0.00226]
Not vertically integrated with a primary school		-0.153***		-0.163		0.0166		0.0207

		[0.0352]		[0.122]		[0.0379]		[0.125]
Number of school units to be managed		-0.0171**		-0.0234		-0.0137*		-0.0201
		[0.00681]		[0.0203]		[0.00728]		[0.0211]
Small town		-0.113***		-0.127		-0.140***		-0.102
		[0.0360]		[0.0879]		[0.0405]		[0.0994]
Province (Ref. Arezzo)								
Pavia		0.0755**		0.0824		-0.0156		0.000564
		[0.0317]		[0.0968]		[0.0355]		[0.0999]
Siracusa		-0.190***		-0.205		-0.288***		-0.231
		[0.0555]		[0.157]		[0.0626]		[0.171]
Constant	-0.168***	0.0246	-0.194***	0.0407	-0.0974***	0.119	-0.0681***	-0.0225
	[0.0143]	[0.0768]	[0.0181]	[0.207]	[0.0140]	[0.0898]	[0.0190]	[0.234]
Random Effects parameters								
var(Constant)				0.0535				0.0557
				[0.0117]				[.012707]
var(Residual)				0.5636				0.6104
				[0.0111]				[.012225]
Observations	5,267	5,265	5,265	5,265	5,069	5,062	5,062	5,062
R-squared	0.294	0.359	0.407		0.296	0.346	0.394	
Adj-r2	0.293	0.357	0.398		0.296	0.343	0.385	
Number of schools				71				68
Log r-likelihood				-6075				-6047

Note: Robust standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1

Table 3: Distribution of schools on the basis of their value added creation in reading and math (absolute values)

		Math				
		Worst performing schools	School performing on the average	Best performing schools	n.a.	Total
Reading	Worst performing schools	7	8	1	0	16
	School performing on the average	10	28	4	0	42
	Best performing schools	0	6	4	3	13
	n.a.	0	0	0	1	1
	Total	17	42	9	4	72

Note: Effectiveness defined as distance from sample mean (5% confidence)

3.3. Absolute *vs* relative school performance

How do absolute and relative performance of schools relate to each other? How much does our adjusted model of cognitive gain create a level playing field across schools?

In Table 4 we report a cross tabulation of the distribution of schools in absolute (raw scores) and relative (value added) terms. We observe that around two thirds of schools are aligned along the main diagonal: their relative performance with respect to sample average does not vary as we move from a static measure to a dynamic one. Our value added estimates do not provide counterintuitive measures of school quality. However, for the remaining third of schools in the sample, estimated value added reveals that their actual performance was either lower (12.7% of cases in reading, 17.6% of cases in math) or higher (18.3% of cases in reading, 16.2% of cases in math) than what could have been surmised by looking at the absolute figures alone. Hence contextual factors and the composition of the student body at the school level play a significant role in final achievements and they are able to enhance or hinder students' learning significantly.

Consequently, value added measures seem to ensure a fairer comparison between schools in the Italian educational system just like highlighted elsewhere over recent decades (OECD, 2008).

Table 4: Cross distribution of schools on the basis of their absolute and relative performance in reading and math (shares)

READING		**Value Added**			Total
		Worst performing schools	*School performing on the average*	*Best performing schools*	
Absolute scores	*Worst performing schools*	0.15	0.10	0.00	0.25
	School performing on the average	0.07	0.38	0.03	0.48
	Best performing schools	0.00	0.11	0.15	0.27
	Total	0.23	0.59	0.18	1.00

MATH		**Value Added**			Total
		Worst performing schools	*School performing on the average*	*Best performing schools*	
Absolute scores	*Worst performing schools*	0.18	0.15	0.01	0.34
	School performing on the average	0.06	0.37	0.03	0.46
	Best performing schools	0.01	0.10	0.09	0.21
	Total	0.25	0.62	0.13	1.00

Note: Effectiveness defined as distance from sample mean (5% confidence) of raw scores and value added scores respectively.

4. Inside the black box: what do schools that add value do?

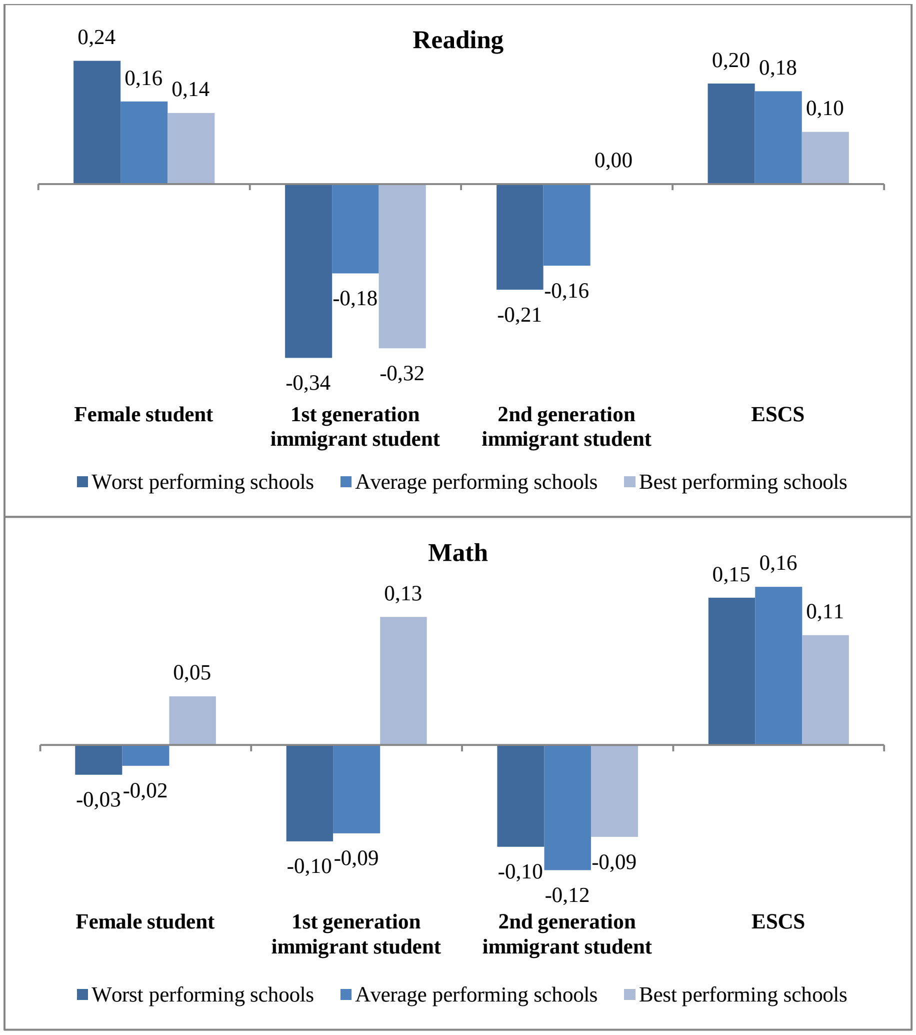
Albeit useful for benchmarking purposes, value-added estimates are usually seen as unable to provide information on what does make a good school. In this section we look at the differences between worst- and best-performing schools in order to gather clues on what is causing their different performances. We also present a three-level variance decomposition of the cognitive gains in reading and math which attempts to quantify the relative weight of teachers' and principal quality in the achievements of students.

4.1. Narrowing the gaps

Full sample estimates like the ones reported in Table 2 provide an indication of each covariate’s average impact on the cognitive progress of pupils. In order to understand what the best schools do differently from the worst ones, we can run sub-sample regressions on schools grouped on the basis of their performances (see results in Appendix, Table A2). We show in Figure 4 how coefficients on individual characteristics vary as we me move from one group to the other.

Best performing schools are able to counterweigh gender gaps: the relative advantage of girls in reading drops as schools become more effective overall and the same is observed for boys in math. The same pattern occurs when we look at the influence of the socio-cultural background on student achievements: the best schools seem to be able to offer more equal opportunities of learning to students coming from disadvantaged families with respect to the worst ones. The picture is more mixed when it comes to the achievement gaps of immigrant students. Best schools are able to narrow the gap of 2nd generation immigrant students in reading but they seem to be as ineffective as the worst ones in math. Conversely, best schools do extremely well with 1st generation immigrant students in math but do as badly as the worst ones in reading.

Figure 4: Impact of individual characteristics on achievements by school effectiveness (estimated coefficients)



Although coefficients related to the foreign origin of students could be imprecisely estimated due to the small sample size, taken at their face value, our results would suggest that best schools tend to be quite selective with their immigrant students, managing to support best those who have better chance to improve on their own. In particular, 2nd generations are brought to the point of mastering the Italian language as well as natives with Italian parents and they achieve good results in reading. On the other hand, 1st generations, largely of Eastern European origin, are given the opportunity to exploit their talent and previous knowledge in math and perform well in that subject. It is worth noticing that the worst-performing schools systematically fail to provide adequate support both in reading and math to immigrant students.

4.2. A variance-decomposition of cognitive gain

By means of a multilevel model that accounts for the nested structure of the data in our sample (students grouped into classes, classes grouped into schools), we can provide a 3-level variance decomposition of the estimated cognitive gain. In Table 5, we report the results obtained with an empty model where *c* indexes over classes within schools:

$$s_{i,c,j}^{m,6} = \alpha_0^{c,j} + \alpha_1 s_{i,c,j}^{m,5} + \alpha_2 \left(s_{i,c,j}^{m,5}\right)^2 + u_{i,c,j}^{m} \qquad (6)$$

and

$$\alpha_0^{c,j} = A + \mu_0^{c,j} + \varepsilon_0^{j} \qquad (6’)$$

More than 80% of the variability lies at the student level in both subjects. The remaining part is split between the class and the school levels in such a way that differences between classes within schools account for a larger part of variability than differences between schools themselves. This is especially true for the cognitive progress in math.

It is interesting to investigate what factors are in operation at each level to obtain some hints on what can be improved in terms of endowments, processes and activities to promote student achievements. We contrast the empty model with an adjusted one that includes all the observable factors at the relevant levels. In particular, we adjust at the class level with the composition of the student group in terms of average socio-cultural background and heterogeneity (ESCS mean and standard error), and at the student and school level with all observable factors as in the models presented in section 3. As we add explanatory variables, all residual variance at different levels can be attributed to the operation of unobserved factors. At the class level, the residual variance approximates the impact of teacher quality (the unobserved factor); similarly, at the school level, the residual variance is attributable to other unobserved factors such as the quality of management.

The variance decomposition obtained with the adjusted model reveals that the quality of management and other unobserved school-level factors account for the largest differences in variability in school performances (76% in reading, 62% in math) but capture a mere 5% of overall variability in cognitive gains. Much more remarkable is the influence of teacher quality on between classes/within schools differences in achievement, which amounts to 77% in reading and 92% in math. The quality of teaching captures up to 7.2% and 11.1% of overall variability in student performance in reading and math respectively.

Table 5: Variance decomposition of cognitive gain (3-levels hierarchical model)

Reading

Levels	Empty model	Adjusted model	Factors
Between students within class	83.2%	4.8%	*Observed individual factors*
		78.4%	*Other unobserved individual factors (residual)*
Between classes within school	9.3%	2.1%	*Class composition*
		7.2%	*Quality of teaching (residual)*
Between schools	7.5%	1.9%	*Observed school characteristics*
		5.7%	*Quality of management + other unobserved factors (residual)*
Total	100.0%	100.0%	

Math

Levels	Empty model	Adjusted model	Factors
Between students within class	80.1%	2.3%	*Observed individual factors*
		77.8%	*Other unobserved individual factors (residual)*
Between classes within school	12.1%	1.0%	*Class composition*
		11.1%	*Quality of teaching (residual)*
Between schools	7.8%	2.9%	*Observed school characteristics*
		4.8%	*Quality of management + other unobserved factors (residual)*
Total	100.0%	100.0%	

5. Conclusions

In this paper we managed to estimate school effects on cognitive gains, based on students' longitudinal test scores, for 72 schools involved in an experiment by the Italian Ministry of Education. Different econometric approaches were used: results seem very robust to different specifications and statistical methodologies. At the end we settled for an adjusted linear model. As the purpose of the exercise is to identify schools which perform best given students' characteristics and the external environment, a lot of effort was devoted to choose the relevant explanatory variables to be included in the regressions. Our guiding principle has been to consider only those variable which are completely outside schools' control. Hence we used as regressors two sets of variables: student characteristics (prior test score, gender, immigrant status, socio-economic background) as Italian schools cannot implement selective admission policies; school context (share of immigrant, disable and disadvantaged students, school complexity, teachers' turnover, geographic location). On the contrary we did not include variables related to the class composition as it belongs to the principal or teachers' domain and it could reflect deliberate choice in terms of ability or social tracking (Ferrer-Esteban, 2011).

In the second part of the paper we tried to look at what makes a good school good. We found out two plausible trails. The first is the ability to narrow the given gaps among students by gender, status and origin. Apparently, good schools are also able to identify the talents of their immigrant students and provide support in specific domains albeit at the expenses of depressing learning in other domains. The second is the quality of teaching which, in our estimates, captures most of the within-school between-classes variability, and overshadows the contribution of the quality of management by the principal on the overall performance of students.

References

Aitkin, M. and N. T. Longford, 1986. Statistical Modelling Issues in School Effectiveness Studies. *Royal Statistical Society*, Series A, 149 (1), 1-43.

Ballou, D., Sanders, W., and Wright, P. (2004). Controlling for student background in valueadded assessment of teachers. *Journal of Educational and Behavioral Statistics, 29* (1), 37-65.

Barbieri, G., Cipollone, P. and P. Sestito, 2007. Labour Market for Teachers: Demographic Characteristics and Allocative Mechanisms. *Giornale degli Economisti*, *GDE (Giornale degli Economisti e Annali di Economia)*, Bocconi University, vol. 66(3), pages 335-373.

Bertola, G. and D. Checchi, 2008. Organizzazione delle risorse scolastiche. Motivazione, organizzazione e carriere degli insegnanti nel sistema pubblico italiano. Working Paper No. 5, Fondazione Giovanni Agnelli, Torino (Italy).

Checchi, D., De Simone, G. and E. Rettore, 2013. Monetary incentives and schools effectiveness. Short term evidence from Italy. *Mimeo*.

Coleman, J., 1966. *Equality of Educational Opportunity*. Washington D.C.: U.S. Department of Health, Education, and Welfare.

Ferrer-Esteban, G., 2011. Beyond the traditional territorial divide in the Italian Education System. Working Paper 43, Fondazione Giovanni Agnelli, Torino.

Fitz-Gibbon, C., 1997. *The Value Added National Project Final Report: Feasibility Studies for a National System of Value-Added Indicators*. London: School Curriculum and Assessment Authority.

Goldstein, H., 1997. Methods in School Effectiveness Research. *School Effectiveness and School Improvement*, 8, 369-95.

Kohn, A., 2000. *The Case Against Standardized Testing: Raising the Scores, Ruining the Schools.* Portsmouth, NH: Heineman.

McCall, M. S., Kingsbury, G. G. and A. Olson, 2004. *Individual Growth and School Success.* Lake Oswego, OR: Northwest Evaluation Association.

Montanaro, P., 2008. Learning divides across the Italian regions: some evidence from national and international surveys. Bank of Italy Questioni di Economia e Finanza (Occasional Papers) 12.

Nichols, S.L. and D.C. Berliner, 2005. *The Inevitable Corruption of Indicators of Educators through High-stakes testing*,Tempe, AZ: Education Policy Reserarch Unit, Arizona State University.

OECD, 2008. Measuring improvements in learning outcomes*: Best practices to assess the value-added of schools*. Paris: OECD.

Reckase, M.D., 2008. *Measurement issues associated with value-added models.* Paper prepared for the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC, November 13-14.

Rothstein, J. (2009). Student Sorting and Bias in Value Added Estimation: Selection on Observables and Unobservables. *Education Finance and Policy*, 4(4).

Rothstein, J., 2010. Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125(1), February 2010, pp. 175-214.

Sanders, W., Saxton, A. and B. Horn, 1997. The Tennessee Value-Added Assessment System: A Quantitative Outcomes-Based Approach to Educational Assessment. In J. M. (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluational Measure?*. Thousand Oaks, CA: Corwin Press, Inc.

Schmidt, W.H., Houang, R.T., and C.C. McKnight, 2005. Value-added research: Right idea but wrong solution? In R. Lissitz (Ed.), *Value-added models in education: Theory and applications*, Maple Grove, MN: JAM Press.

Tong, Y., and Kolen, M.J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20*(2), 227-253.

Willms, J.,and Raudenbush, S. (1989, 26(3)). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, 209-232.

Young, D. (1998). Rural and urban differences in student achievement in science and mathematics: A multilevel analysis. *School Effectiveness and School Improvement*, 9, 386-418.

Young, M.J. (2006). Vertical scales. In S.M. Downing and T.M. Haladyna (Eds.), *Handbook of test development* (pp. 469-485). Mahwah, NJ: Lawrence Erlbaum Associates.

Appendix

Table A1: School characteristics and the probability to participate in the pilot program – Probit model

VARIABLES	Marginal effects
Province = Arezzo	-0.481***
	[0.167]
Province = Pavia	-0.0103
	[0.248]
School in Province Capital	0.237*
	[0.127]
Total number of students	-0.000983
	[0.000773]
(Inflow – Outflow) of students during the school year	-2.079
	[3.695]
Student drop out rate	-8.773
	[6.498]
Grade retention rate	0.395
	[1.336]
Final exam failure rate	5.726
	[5.461]
Average size of classes in the first year	-0.0544**
	[0.0241]
Share of classes with full-day schooling	-0.306
	[0.276]
Students/PC (desktop&notebook) ratio	-0.000103
	[0.000931]
Students/Interactive Whiteboard ratio	-0.000819
	[0.000553]
Average age of teachers	-0.0176
	[0.0325]
Share of male teachers	0.531
	[0.623]
Share of teachers with a permanent contract	-0.791
	[0.802]
Share of teachers for students with special needs	-0.323
	[0.851]
Pupil/Teacher ratio	0.00168
	[0.0516]
Pupil/Non-teaching staff ratio	0.00128
	[0.0110]
Average working days missing for non-teaching staff ratio	-0.0370
	[0.0688]
Average working days missing for teachers	-0.247**
	[0.111]
Share of female students	0.00695
	[0.00777]
Share of students born abroad (1st generation)	0.0181
	[0.0152]
Share of native students with parents born abroad (2nd generation)	0.0151
	[0.0186]
Average Socio-Economic Status of students	0.134
	[0.195]

Observations	115
Pseudo-r2	0.231

Note: Standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1

Source: Checchi-De Simone-Rettore (2013).

Table A2: Impact of individual characteristics on achievements - Subsample estimates

	Reading			**Maths**		
Dep. Variable: Test score at grade 6	*Worst performing schools*	*Average performing schools*	*Best performing schools*	*Worst performing schools*	*Average performing schools*	*Best performing schools*
Test score at grade 5	0.628***	0.607***	0.591***	0.536***	0.581***	0.609***
	[0.0316]	[0.0195]	[0.0318]	[0.0322]	[0.0175]	[0.0404]
Test score at grade 5 - squared	0.200***	0.222***	0.231***	0.177***	0.199***	0.159***
	[0.0219]	[0.0125]	[0.0197]	[0.0213]	[0.0139]	[0.0267]
Female student	0.244***	0.159***	0.140***	-0.0303	-0.0226	0.0471
	[0.0456]	[0.0276]	[0.0444]	[0.0476]	[0.0278]	[0.0636]
Immigrant status (Ref. Native with Italian parents)						
1st generation immigrant student	-0.234**	-0.0891	-0.346***	-0.0615	-0.0374	0.159
	[0.113]	[0.0726]	[0.125]	[0.117]	[0.0610]	[0.144]
2nd generation immigrant student	-0.193*	-0.149*	0.0883	-0.0974	-0.114*	-0.0909
	[0.117]	[0.0854]	[0.102]	[0.0879]	[0.0678]	[0.148]
Grade repeter in primary school	-0.246*	-0.240***	0.0985	-0.114	-0.136**	-0.130
	[0.133]	[0.0780]	[0.120]	[0.141]	[0.0621]	[0.165]
ESCS	0.197***	0.178***	0.103***	0.145***	0.157***	0.106***
	[0.0240]	[0.0162]	[0.0250]	[0.0301]	[0.0152]	[0.0316]
Constant	-0.500***	-0.155***	0.0581	-0.313***	-0.0680***	0.283***
	[0.0391]	[0.0235]	[0.0400]	[0.0414]	[0.0236]	[0.0500]
Observations	1,238	3,020	1,007	1,073	3,268	721
R-squared	0.391	0.393	0.376	0.356	0.382	0.374
Adj-r2	0.380	0.384	0.364	0.342	0.373	0.361

Note: Robust standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1