



Bivariate multilevel models for the analysis of mathematics and reading pupils' achievements

C. Masci, F. Ieva, T. Agasisti & A. M. Paganoni

To cite this article: C. Masci, F. Ieva, T. Agasisti & A. M. Paganoni (2016): Bivariate multilevel models for the analysis of mathematics and reading pupils' achievements, Journal of Applied Statistics, DOI: [10.1080/02664763.2016.1201799](https://doi.org/10.1080/02664763.2016.1201799)

To link to this article: <http://dx.doi.org/10.1080/02664763.2016.1201799>



Published online: 29 Jun 2016.



Submit your article to this journal [↗](#)




View related articles [↗](#)



View Crossmark data [↗](#)



Bivariate multilevel models for the analysis of mathematics and reading pupils' achievements

C. Masci^a, F. Ieva^b , T. Agasisti^c and A. M. Paganoni^a

^aDepartment of Mathematics, Politecnico di Milano, Milano, Italy; ^bDepartment of Mathematics 'Federigo Enriques', Università Statale di Milano, Milano, Italy; ^cDepartment of Management, Economics and Industrial Engineering, Politecnico di Milano, Milano, Italy

ABSTRACT

The purpose of this paper is to identify a relationship between pupils' mathematics and reading test scores and the characteristics of students themselves, stratifying for classes, schools and geographical areas. The data set of interest contains detailed information about more than 500,000 students at the first year of junior secondary school in the year 2012/2013, provided by the Italian Institute for the Evaluation of Educational System. The innovation of this work is in the use of multivariate multilevel models, in which the outcome is bivariate: reading and mathematics achievement. Using the bivariate outcome enables researchers to analyze the correlations between achievement levels in the two fields and to predict statistically significant school and class effects after adjusting for pupil's characteristics. The statistical model employed here explicates account for the potential covariance between the two topics, and at the same time it allows the school effect to vary among them. The results show that while for most cases the direction of school's effect is coherent for reading and mathematics (i.e. positive/negative), there are cases where internal school factors lead to different performances in the two fields.

ARTICLE HISTORY

Received 20 May 2015
Accepted 12 June 2016

KEYWORDS

Pupils' achievement;
multilevel models; bivariate
models; school and class
effects; value-added

1. Introduction and motivation

Nowadays, the analysis of the differences in educational attainments between groups of students and across schools and classes is becoming increasingly interesting. Due to the increasing demand for public education accountability, studies on this topic are carried out in order to test and improve the educational system and to understand which variables mostly affect it (see [4,9,18]).

In Italy, the Italian Institute for the Evaluation of Educational System (hereafter INVALSI, founded in 2007) assesses students abilities in reading and mathematics at different stages. This happens at the end of the second and fifth year of primary school (when pupils are aged 7 and 10, respectively), at the end of the first and third year of lower secondary school (aged 11 and 13) and at the end of the second year of upper secondary school (aged 15). Students are requested to answer questions with both multiple choices

and open-ended questions, that test their ability in reading and mathematics. This is a way to test knowledge and reasoning that pupils should have learned in their school career and the results are used to evaluate students, but primarily schools and classes. Also, they are requested to compile a questionnaire about themselves, their family, their parents' educational level and their socio-economic situation.

The institutional organization of the Italian educational system is based on the strong assumptions about its equality purposes, among which all schools and classes provide similar educational standards. Therefore, recent aggregate data provided by INVALSI show that this is not the case, and that a significant portion of variance in students' test scores is attributable to the structural between-schools differences (see [1,12]). The use of multilevel models in this field have enabled researchers to catch the stratification between schools, but, up to this time, only in univariate cases – that is, studying the results of one single topic (mathematics or reading).

Several studies are present in literature on the mathematics and reading achievements of Italian students, where they are treated as separate, applying univariate multilevel linear models (see [5,6,14,15]) to analyze how the outcome variables (mathematics or reading achievement) depend on the students' characteristics, and which are the schools/classes affecting students' achievements most (see [1]). Regarding the characteristics of students' profile, it emerged that, for instance, first and second generation immigrant students obtain, on average, worse test scores than native Italian students, both in reading and mathematics; females have better average results than males in reading, but worse ones in mathematics; early/late-enrolled students have worse performances than 'regular' students. Also, big differences elapse between North, Center and South of Italy: students in the South have on average lower results than students in the North and the aspects of students' profile weigh in a different way across the geographical macro-areas, emphasizing the need to have three different models to explain the completely different phenomena. Furthermore, the results asserted that the choice of the school and especially of the class can influence the students' performances, acting beyond the pure effect of their socio-economic background. In previous studies (see [1,12]) emerged some common behaviors about students' characteristics and we noted a good correlation between the reading and mathematics achievements. As we explained before, there are aspects of student's profile that weigh in the same way both in reading and mathematics. At this point, we held that could be interesting to analyze the relationships and the interactions between the school/class effects in reading and mathematics. So that, the main purpose of this work is to fit multilevel linear models, for the bivariate outcome of reading and mathematics INVALSI achievements. In particular, these two outcomes are regressed against a series of variables at student's level with the values-added of schools and classes (random effects). The use of a bivariate model for studying the determinants of students' test scores is innovative not only for an empirical reason, or for the use of the statistical modeling, but also for an important conceptual reason. When following the traditional literature in the economics field, the estimation of an educational production function considers the impact of several inputs (students, schools and contextual variables) on a measure of output (see the discussion in [8]). Of course, several different models can be estimated empirically, by alternating the output under scrutiny to check whether the factors are affecting in equal way the test scores in reading and mathematics, for example. Therefore, such an empirical procedure assumes that the production process of learning can be described as separating the role of each input

for a specific output or, in other words, without paying attention to the joint production of the two outputs. With an adequate statistical modeling of bivariate outcomes, instead, the production of learning is described as jointly operating on both outputs simultaneously, so that we can estimate the role of each input (variable) in producing not reading or math, but reading and math learning. It is important to take into account that the aim of this paper is to apply multilevel models for estimating how much of the variance of students' test scores is attributable to structural differences between schools and classes. We refer to this as 'school/class effect'. In other words, we aim at capturing how attending a specific school/class statistically influences the students' test scores, after accounting for their individual characteristics. We do not investigate mechanisms through which such school/class effect can be considered as causal, instead we focus on statistical differences in test scores between schools - thus, the term 'school effect' should be interpreted accordingly. Concerning this, we remind that multilevel models may have two different interpretation, that is, they may be seen as: (i) 'structural models', in the sense that the regression coefficients are causal parameters and the error terms represent the effects of omitted covariates or (ii) 'statistical models', where the coefficients represent associations or linear projections and the error terms are uncorrelated with all covariates by definition (see [3]). In this work, we consider our models as 'statistical models'.

The steps that we carry out in this paper are: (a) to examine the associations between pupils' characteristics, such as profile, socio-cultural background, household, cultural resources, and pupil's achievement, (b) to discover how the school and class effects positively or negatively influence specific types of students' profile and how they are correlated when considering the joint 'production' of reading and mathematics achievements. We will figure out that there are big differences both between school and class effects and between mathematics and reading, showing discrepancies between and within schools and across the country. In particular, we will discover that the class effects weigh more than the school ones, proving that the choice of the class may influence the students' performances more than the choice of the school and that the school/class effects in mathematics and reading may be influenced by different aspects. Furthermore, while the school effects in the two topics will prove to be pretty coherent, identifying good/bad schools, the two class effects will not, proving to be uncorrelated. Lastly, we will figure out that the school effect is independent of the contained class effects, suggesting that good (or bad) schools do not necessarily contain good (or bad) classes.

The paper is organized as follows: in Section 2 we present the data set and the models that we use; in Section 3 we analyze students nested in schools by means of bivariate two-level linear models, stratifying the models by macro-areas; in Section 4 we analyze students nested in classes, which in turn nested in schools by means of a three-level model; Section 5 contains discussion and conclusions. All the bivariate multilevel models are fitted using the software ASReml [7] and the other models and analysis are carried out in the statistical software R [16].

2. Data set and models

In this section we present the data set of interest and the statistical tools requested to make the analysis.

Our resources are two separate set of data, containing information about more than 500,000 students attending the first year of junior secondary school in the year 2012/2013, provided by INVALSI. The former contains the mathematics achievements and the latter the reading ones, followed by the corresponding information about students, classes and schools. The reading data set contains information about 510,933 students and the mathematics one about 509,371. We linked these two data set, retaining only the students that have both the test scores of mathematics and reading, followed by all the variables presented in the two set of data. The deterministic linkage of the two data set is possible thanks to the anonymous student ID that is known for each pupil. We then obtain a new data set containing 507,229 students, whose both the achievements in mathematics and reading are known, and 50 related variables, losing very few individuals.

At pupil's level, the following information is available: gender, immigrant status (Italian, first generation, second generation immigrant), if the student is early-enrolled (i.e. was enrolled for the first time when five years-old, the norm being to start the school when six years-old), or if the student is late-enrolled (this is the case when the student must repeat one grade, or if he/she is admitted at school one year later if immigrant). The data set contains also information about the family's background: if the student lives or not with both parents (i.e. the parents are died, or are separated/divorced), and if the student has siblings or not. Lastly, INVALSI collects information about the socioeconomic status of the student, by deriving an indicator (called ESCS-Economic and Social Cultural Status), which is built in accordance to the one proposed in the OECD-PISA framework. In other words, it is built considering (i) parents' occupation and educational titles, and (ii) possession of certain goods at home (for instance, the number of books). Once measured, this indicator has been standardized to have mean zero and variance one. The minimum and maximum observed values in the INVALSI data set are -3.11 and 2.67 , respectively. In general, pupils with ESCS equal to or greater than 2 are very socially and culturally advantaged (high family's socioeconomic background). Among data, there are also the INVALSI scores in the Mathematics and Reading tests at grade 5 of the previous year (ranging between 0 and 100), which are used as a control in the multilevel model to specify a Value-Added estimate of the school's fixed effect. It is well known from the literature that education is a cumulative process, where achievement in the period t exerts an effect on results of the period $t+1$. The data set also allows us to explore several characteristics at class level, among which the class-level average of several individuals' characteristics (e.g. class-average ESCS, the proportion of immigrant students, etc.). Of particular importance, there is a dummy for schools that use a particular schedule for lessons ('Tempo Pieno', classes comprise educational activities in the afternoon, and no lessons on Saturday, while traditional classes end at lunchtime, from Monday to Saturday). Also the variables at school level measure some school-average characteristics of students, such as the proportion of immigrants, early and late-enrolled students, etc. Two dummies are included to distinguish (i) private schools from public ones, and (ii) 'Istituti Comprensivi', which are schools that include both primary and lower-secondary schools in the same building/structure. This latest variable is relevant to understand if the 'continuity' of the same educational environment affects (positively or negatively) students results. Some variables about size (number of students per class, average size of classes, number of students of the school) are also included to take size effects into account. Lastly, regarding geographical location, we include two dummies for schools located in Central and Southern Italy; some previous literature, indeed, pointed

at demonstrating that students attending the schools located in Northern Italy tend to have higher achievement scores than their counterparts in other regions, all else equal (see [1,12]). As we have the anonymous student ID, we have also the encrypted school and class IDs that allow us to identify and distinguish schools and classes. The outputs MS and RS (hereafter, respectively, the score in the Mathematics and Reading standardized test administered by INVALSI) are expressed as ‘cheating-corrected’ scores (CMS and CRS). In fact, INVALSI estimates the propensity-to-cheating as a percentage, based on the variability of intra-class percentage of correct answers, modes of wrong answers, etc.; the resulting estimates are used to ‘deflate’ the raw scores in the test. These variables take values between 0 and 100.

Unfortunately, there are lots of missing data in the score at grade 5, both in mathematics and reading achievements. This kind of data may have been lost in the passage of information between primary and junior secondary schools. Since having longitudinal data is very important for this study, we omit the individuals with missing data at grade 5, losing almost 300,000 students. An other possibility is to impute the missing data: there are different techniques to do that, but from an analysis on this topic, it emerges that the missing data are not completely at random, so we can not impute them in a correct way. That is why we decide to delete the individuals with missing data and to create a reduced data set in which we consider only the individuals with all the variables available. In order to be aware of the potential bias induced by this operation, we assess the representativeness of this subsample in Section 2.1. The final and reduced data set collects 221,529 students, almost half of the initial data set, within 16,246 classes, within 3920 schools.

Hereafter, all the analysis are made on this reduced data set. The variables and some related descriptive statistics are presented in Table 1.

2.1. Representativeness of the sub-sample

As mentioned before, in the reduced database we mainly discard all the statistical units for which the 5th year of primary school math and reading score (CMS5 and CRS5) is missing. It is worth evaluating from a statistical point of view the representativeness of this sub-sample with respect to the entire population. Since the sample size at pupil’s level is extremely high, is quite impossible to find a non-significant difference in statistics summarizing the student’s level features. Moreover the CMS5 and CRS5 scores are data transmitted to INVALSI at school level. For these reasons we check the representativeness of the sub-sample studying the distributions of the school’s level variables. For the continuous ones we perform a non-parametric comparison test (Wilcoxon test) to detect possible differences in the stochastic distributions generating data. For the dichotomic ones we perform a comparison test between proportions.

In particular, the mean ESCS in the reduced data set seems to be quite higher than in the original data set ($p\text{-value} = 3.108 * 10^{-05}$), as well as the percentages of both first and second generation immigrant students ($p\text{-value} = 0.0024$ and 0.0012 , respectively). There is no statistical evidence for difference about the percentage of females ($p\text{-value} = 0.1848$), late-enrolled students percentage ($p\text{-value} = 0.863$) and early-enrolled students percentage ($p\text{-value} = 0.109$). There is no statistical evidence for difference about the proportion of private and public schools ($p\text{-value} = 0.4782$). Regarding the geographical areas, the Center is well represented ($p\text{-value} = 0.7364$), while the North results to be

Table 1. Variables of the database.

Level	Type	Variable name	Mean	sd
Student	–	Student ID	–	–
Student	(Y/N)	Female	49.8%	–
Student	(Y/N)	First generation immigrants	4.4%	–
Student	(Y/N)	Second generation immigrants	4.9%	–
Student	num	ESCS	0.24	1.02
Student	(Y/N)	Early-enrolled student	1.6%	–
Student	(Y/N)	Late-enrolled student	2.8%	–
Student	(Y/N)	Not living with both parents	12.6%	–
Student	(Y/N)	Student with siblings	83.3%	–
Student	%	Cheating	0.016	0.05
Student	num	Written reading score	9.41	2.74
Student	num	Oral reading score	6.80	1.13
Student	num	Written mathematics score	9.48	2.75
Student	num	Oral mathematics score	6.88	1.35
Student	num	CMS5-5th year Primary school mathematics score	70.5	16.30
Student	num	CRS5-5th year Primary school reading score	74.5	13.50
Class	–	Class ID	–	–
Class	num	Mean ESCS	0.18	0.48
Class	%	Female percentage	43.7	10.07
Class	%	First generation immigrant percent	5.4	6.47
Class	%	Second generation immigrant percent	4.7	5.83
Class	%	Early-enrolled student percent	1.4	3.24
Class	%	Late-enrolled student percent	6.2	6.11
Class	%	Disable percentage	5.8	5.58
Class	count	Number of students	23	3.49
Class	(Y/N)	'Tempo pieno'	0.023%	–
School	–	School ID	–	–
School	num	Mean ESCS	0.18	0.41
School	%	Female percentage	43.3	5.46
School	%	First generation immigrant percent	5.4	4.65
School	%	Second generation immigrant percent	4.6	4.06
School	%	Early-enrolled student percent	1.5	2.23
School	%	Late-enrolled student percent	6.3	3.94
School	count	Number of students	143	76.52
School	count	Average number of students	22.6	2.94
School	count	Number of classes	6.2	3.05
School	(Y/N)	North	52%	–
School	(Y/N)	Center	18%	–
School	(Y/N)	South	30%	–
School	–	District	–	–
School	(Y/N)	Private	3.1%	–
School	(Y/N)	'Istituto comprensivo'	65.8%	–
Outcome	num	CMS-Mathematics Score corrected for Cheating	47.4	17.67
Outcome	num	CRS-Reading Score corrected for Cheating	65	14.65

over represented ($p\text{-value} = 9.927 * 10^{-05}$) and the South results to be under represented ($p\text{-value} = 2.244 * 10^{-05}$). This suggests a higher efficiency in the transmission of administrative information to INVALSI and between schools in the North than in the South. Being this behavior voluntarily driven or not, it represents a problem for the evaluation of the national educational system. Overall, the reduced data set used in this paper is substantially representative of the original population, with the only exception of the proportion of schools in the South and in the North. Albeit the use of the reduced sample can be criticized on this ground, it has two major advantages that justify our choice. First, the performance at grade 5 is strongly predictive of the test score at grade 6, and the dismissal of such important control can generate a problem of omitted variables that is statistically more serious than the problem of macro-areas' representativeness. Second, the inclusion

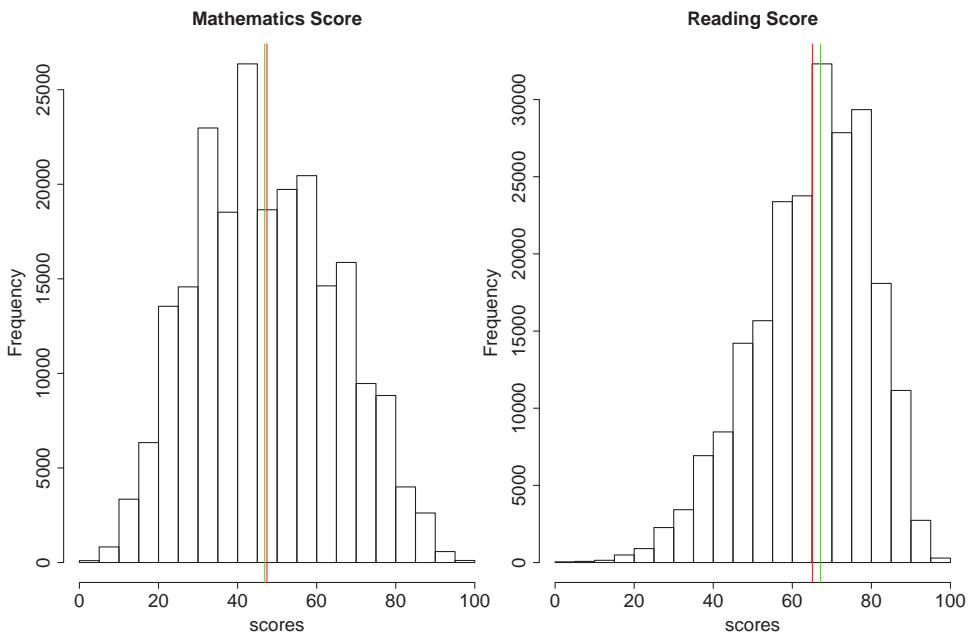


Figure 1. Histogram of Corrected Reading and Mathematics Score of pupils in the Invalsi database. The red lines refer to the mean, the green ones to the median.

of prior achievement allows considering the analysis in a VA (Value added) fashion; as a consequence, the effects of attending a specific j th school is not predicted through a simple cross-sectional variation, but also with reference to a longitudinal variation in relative test scores.

The main statistical tools employed in the analysis are bivariate multilevel linear models. We first develop a two-level model in which students (level 1) are nested in schools (level 2). We consider only variables at student level with random effects on schools. This allows us to individuate the relationship between the test results (answer variable) and the characteristics of student's profile (covariates at level 1) and to predict the random effects (school effects). In a second step, by means of simple linear regression, we look for the variables at school level that are correlated with the predicted random effects. Furthermore, we later introduce a third step of the analysis in which we add the class level, considering a three-level linear model in which students (level 1) are nested in classes (level 2) that are in turn nested in schools (level 3). In this case, in addition to individuate the association between test results (answer variable) and characteristics of student's profile, we predict two random effects: classes and schools.

As we introduced before, the outcome variable of these models is bivariate: reading and mathematics achievements. Figure 1 shows the histograms of the two test results.

It is immediately clear that the two distributions are different and that the CRSs are on average higher than the CMSs. There is a positive correlation between the performances of students in the two topics, CMS and CRS. By a test of correlation (Pearson's product moment), we obtain a correlation coefficient of 0.59 with a high significance (p -value less than 2.2×10^{-16}).

3. Bivariate two-level linear model: students nested in schools

The first model that we fit is a bivariate two-level linear model in which pupils (level 1) are nested in schools (level 2). Let N be the total number of pupils and n_j the total number of pupils in the school j , for $j = 1, \dots, J$, such that $\sum_{j=1}^J n_j = N$. For each school $j = 1, \dots, J$ and pupil $i = 1, \dots, n_j$:

$$\mathbf{y}_{ij} = \boldsymbol{\beta}_0 + \sum_{k=1}^K \boldsymbol{\beta}_k x_{kij} + \mathbf{b}_j + \boldsymbol{\epsilon}_{ij}, \quad (1)$$

where \mathbf{y}_{ij} is the bivariate outcome that represents the mathematics and reading achievements of pupil i in school j ; $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \dots, \boldsymbol{\beta}_K)$ is the bivariate $(K + 1)$ -dimensional vector of parameters; x_{kij} is the value of the k th predictor variable at student's level; $\mathbf{b} \sim N_2(\mathbf{0}, \Sigma)$ is the matrix of the bivariate random effects (mathematics and reading) at school level and $\boldsymbol{\epsilon} \sim N_2(\mathbf{0}, W)$ is the matrix of errors, where Σ and W are, respectively, the variance/covariance matrices of the random effects and of the errors.

We assume \mathbf{b} independent of $\boldsymbol{\epsilon}$. Furthermore, the usual assumptions are made about the random effects at school level: \mathbf{b}_j for the j th school is assumed to be Gaussian distributed independent of any predictor variables that are included in the model and uncorrelated across the levels.

Using the software *ASReml*, we can fit this bivariate model and obtain the estimates of the coefficients and of the random effects, identifying how the characteristics of pupils are associated with their achievements and computing the values-added given by the schools. Table 2 shows the results. Note that we managed to use the CRS5/CMS5 as a regressor only for the reading/mathematics achievement, respectively, because the reading achievement does not depend on the mathematics score at grade 5 and viceversa.

Now, we can analyze the relationships between the test scores and the characteristics of students, comparing the estimates of the coefficients of the two topics. As we anticipated before, the coefficients of the variable 'female' of the two topics are almost opposites: being a female is positively associated with better results in reading, but with worse ones in mathematics, suggesting that on average males are better in mathematics, while females are better in reading. The coefficients reveal that, as expected, immigrants perform much worse in reading than their Italian counterparts, and especially first generation immigrants also have (slightly) lower than second generation immigrants for whom, however, significant gap remains. This would suggest that being born in Italy is not sufficient to close the gap in Italian language proficiency, if it is not practiced in the family life. Immigrants also show lower scores in mathematics, and it is strange to notice how the gap between first and second generation immigrants is more favorable for the former; this can be interpreted as whether the language is not a key barrier for the proficiency in mathematics, once that the specific differences in reading's score are taken into account (remember, indeed, that the correlation of immigrant status with performance is estimated in a bivariate setting). Being a student in the Center and especially in the South of Italy is associated with worse results in both the fields, especially in mathematics: students of the South have worse results than students of the North. Being early/late-enrolled weighs negatively in both the fields. The ESCS and the score at grade 5 are positively associated with the achievements and have similar coefficients in both the fields. Regarding the ESCS, this means that the socio-economic status of students and the family background weigh positively on the performances: students

Table 2. ML estimates of model (1) fitted to the data set.

Fixed effects	Mathematics coeff (se)	Reading coeff (se)
Intercept	14.91*** (0.201)	30.44*** (0.187)
Female	−2.211*** (0.057)	2.134*** (0.049)
First generation immigrant	−1.511** (0.159)	−3.921*** (0.137)
Second generation immigrant	−2.281*** (0.137)	−3.548*** (0.117)
South	−6.437*** (0.211)	−4.670*** (0.167)
Center	−2.699*** (0.262)	−1.163*** (0.207)
Early-enrolled student	−0.793** (0.229)	−0.792** (0.196)
Late-enrolled student	−2.744*** (0.192)	−3.638*** (0.164)
ESCS	2.625*** (0.031)	2.211*** (0.027)
Not living with both parents	−1.463*** (0.087)	−1.104*** (0.074)
Student with siblings	0.049 (0.078)	−0.644*** (0.067)
CS5	0.505*** (0.002)	0.476*** (0.002)
Variance/covariance matrix of random effects	$\begin{pmatrix} 23.04 & 5.51 \\ 5.51 & 13.08 \end{pmatrix}$	
Variance/covariance matrix of error	$\begin{pmatrix} 180.5 & 63.13 \\ 63.13 & 132.25 \end{pmatrix}$	
Size		
Number of observations	221,529	
Number of groups (school)	3920	

Notes: Asterisks denote different levels of significance.

** $0.0001 \leq p\text{-val} \leq 0.001$.

*** $p\text{-val} \leq 0.0001$.

with high value of ESCS have better performances. Lastly, the positive influence of the score at grade 5 suggests that there is a continuity in the students' efficiency.

Looking at the variance/covariance matrix of the random effect, it is evident that the variability of the mathematics random effect is much higher than the reading one (23.04 vs. 13.08), therefore, attending a specific school influences more the results in mathematics than in reading. The two effects are positively correlated (0.307). Figure 2 shows the variability of the marginal random effects.

The school effects in mathematics are more scattered than the reading ones, that is they are farer from zero. This means that the effect that schools give in mathematics is stronger than the reading one: the school influences the mathematics achievements more than the reading ones. To test this difference in variability, we implement a non-parametric Levene's test. We obtain a p -value less than 2.2×10^{-16} , proving that the variances of the random effects of the two topics are different.

3.1. Differences across macro-areas

From previous results and previous studies (see [1,12]) emerges that big discrepancies elapse between the three geographical macro-areas and the variables that result to be

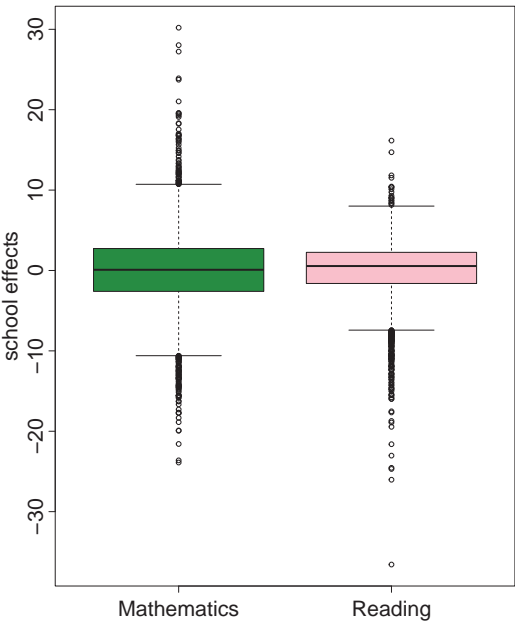


Figure 2. Predicted school effects \hat{b}_j in mathematics and reading.

Table 3. Number of students, classes and schools in the three geographical areas.

Size	North	Center	South
Students	115,368	39,847	66,314
Classes	7754	3066	5426
Schools	1800	688	1432

significant across them are different. Indeed, they may be considered as three different educational systems. In particular, since we assume that the error and the random effects are uncorrelated with the covariates, the geographical variable and differences in influential variables across macro-areas may violate this assumption. Therefore, starting from now, we fit the next models for each macro-area – so that, dividing the data set by macro-area and implementing three different models –, in order to avoid correlation problems and to point out the differences across the areas and the trends inside them.

Before starting with the models, Table 3 shows the size of the data, counting how many students, classes and schools are in each macro-area.

We start with the two-level mixed model (1) fitted for each macro-area:

$$\mathbf{y}_{ij}^{(R)} = \boldsymbol{\beta}_0^{(R)} + \sum_{k=1}^K \boldsymbol{\beta}_k^{(R)} x_{kij} + \mathbf{b}_j^{(R)} + \boldsymbol{\epsilon}_{ij}^{(R)}, \tag{2}$$

where $R = \{\text{North}, \text{Center}, \text{South}\}$. The estimates of model (2) for each macro-area are reported in Table 4.

Looking at the estimates of the three models, we observe that, in general, the coefficients of variables immigrants and early-enrolled students of the South are closer to zero

Table 4. ML estimates of model (2) fitted for each macro-area.

Fixed effects	North math	Center math	South math
Mathematics	(se)	(se)	(se)
Intercept	6.24*** (0.21)	12.83*** (0.39)	20.63*** (0.36)
Female	−1.82*** (0.07)	−2.87*** (0.14)	−2.15*** (0.11)
First generation imm	−1.28** (0.18)	−1.17** (0.35)	0.18 (0.53)
Second generation imm	−2.32*** (0.15)	−1.44** (0.29)	−0.55 (0.45)
Early-enrolled student	−2.33*** (0.44)	−0.51 (0.59)	−0.24 (0.31)
Late-enrolled student	−2.71*** (0.23)	−1.71*** (0.43)	−3.51*** (0.46)
ESCS	2.10*** (0.04)	2.56*** (0.07)	3.28*** (0.06)
Not living with both parents	−1.37*** (0.11)	−1.50*** (0.19)	−1.57*** (0.19)
Student with siblings	0.15 (0.09)	−0.10 (0.17)	−0.03 (0.18)
CMS5	0.62*** (0.01)	0.50*** (0.01)	0.33*** (0.01)
Fixed effects	North read	Center read	South read
Reading	(se)	(se)	(se)
Intercept	24.12*** (0.22)	31.16*** (0.39)	33.61*** (0.33)
Female	2.16*** (0.06)	1.89*** (0.12)	2.21*** (0.09)
First generation imm	−3.93*** (0.15)	−3.75*** (0.30)	−1.66** (0.46)
Second generation imm	−3.74*** (0.13)	−3.22*** (0.25)	−1.16** (0.39)
Early-enrolled student	−2.04*** (0.37)	−0.81 (0.51)	−0.37 (0.27)
Late-enrolled student	−3.42*** (0.19)	−2.85*** (0.37)	−4.92*** (0.41)
ESCS	1.76*** (0.03)	2.21*** (0.06)	2.80*** (0.05)
Not living with both parents	−1.01*** (0.09)	−1.44*** (0.17)	−1.05*** (0.17)
Student with siblings	−0.55*** (0.08)	−0.63*** (0.15)	−0.74*** (0.15)
CRS5	0.56*** (0.01)	0.45*** (0.01)	0.36*** (0.01)
	North	Center	South
Variance/covariance matrix of random effects	$\begin{pmatrix} 9.74 & 1.85 \\ 1.85 & 11.2 \end{pmatrix}$	$\begin{pmatrix} 14.8 & 5.31 \\ 5.31 & 12.7 \end{pmatrix}$	$\begin{pmatrix} 43.6 & 8.36 \\ 8.36 & 15.7 \end{pmatrix}$
Variance/covariance matrix of residuals	$\begin{pmatrix} 154 & 47 \\ 47 & 113 \end{pmatrix}$	$\begin{pmatrix} 182 & 64 \\ 64 & 159.6 \end{pmatrix}$	$\begin{pmatrix} 210 & 82 \\ 82 & 15 \end{pmatrix}$
Size			
Number of observations	115,368	39,847	66,314
Number of groups (school)	1800	688	1432

Notes: Asterisks denote different levels of significance.

**0.0001 ≤ *p*-val ≤ 0.001.

****p*-val ≤ 0.0001.

than those of the North. Particularly, for immigrant students, this can be explained by the high presence of immigrants in the North respect to the South. The ESCS, instead, has greater coefficients in the South than in the North (3.28 and 2.8 against 2.1 and 1.7), suggesting that in the South, the socio-cultural back-ground is very important in the students' achievements: pupils with high ESCS have better results. Lastly, the score at grade 5 weighs more in the North than in the South in both the topics (0.62 and 0.56 against 0.33 and 0.36), emphasizing a greater continuity in student performances.

Let's now look at the variance/covariance matrices of the random effects. The three matrices seem quite different: instead of the North and the Center, where the variances of the random effects of mathematics and reading are almost the same (respectively, 9.74 vs. 11.2 and 14.8 vs. 12.7), in the South the variance of the random effects of mathematics is much higher than the reading one (43.6 vs. 15.7), meaning that the school weighs more in mathematics than in reading. The correlation coefficients between the two vectors of random effects are, respectively, 0.17 in the North, 0.39 in the Center and 0.32 in the South.

3.1.1. Statistical test for the comparison of two or more independent linear models' coefficients

In order to sustain the conclusions regarding the differences between North, Center and South of Italy in terms of students' characteristics, we introduce a test to compare the coefficients of two (or more) independent linear models. The aim is to create a way to compare the coefficients of different models and a test that allows us to assert whether they are different or not. Let's take into account two simple linear regression models, like

$$\text{Model A : } y = \alpha_0 + \alpha_1 x + \epsilon, \quad (3)$$

$$\text{Model B : } \tilde{y} = \beta_0 + \beta_1 z + \tilde{\epsilon} \quad (4)$$

with the usual normality hypothesis. We consider that

$$\hat{\alpha}_1 \sim N(\alpha_1, \sigma_{\alpha_1}^2), \quad (5)$$

$$\hat{\beta}_1 \sim N(\beta_1, \sigma_{\beta_1}^2), \quad (6)$$

$$\hat{\beta}_1 \perp \hat{\alpha}_1. \quad (7)$$

We generate M sub-samples from the data used in model A and other M sub-samples from the data used in model B and we estimate the coefficients of both the M models, obtaining two samples:

$$\hat{\alpha}_1^1, \dots, \hat{\alpha}_1^M \text{ i.i.d. } \sim N(\alpha_1, \sigma_{\alpha_1}^2). \quad (8)$$

and

$$\hat{\beta}_1^1, \dots, \hat{\beta}_1^M \text{ i.i.d. } \sim N(\beta_1, \sigma_{\beta_1}^2). \quad (9)$$

Now, if we want to test

$$H_0 : \beta_1 = \alpha_1 \quad \text{vs.} \quad H_1 : \beta_1 \neq \alpha_1 \quad (10)$$

we compute a *t*-test for the means of these two populations that have different and unknown variances. In the same way, if we have multiple regressions, instead of using a

t -test, we will use an Hotelling test comparing the means of Gaussian vectors. Moreover, if we want to compare vectors of coefficients of more than two models, we can use a MANOVA. In our case, we would like to compare the coefficients of three models: North, Center and South of Italy. We fit three simple regression models, one for each macro-area, only using the variables at student level seen before:

$$y^{(R)} = \beta_0^{(R)} + \sum_{k=1}^K \beta_k^{(R)} x_k + \epsilon^{(R)}. \quad (11)$$

We create $M = 1,000$ sub-samples of size 1,000 for each macro-area and we compute a MANOVA to compare the means of the three vectors of coefficients. The p -value of the MANOVA results to be less than $2.2 * 10^{-16}$, so that, we can assert that there is a statistically significant difference between the coefficients of the three macro-areas: the variables at student's level are associated with the test results in different ways across the three macro-areas.

3.1.2. Comparing variance matrices

To test if there is really a significant difference between the three variance/covariance matrices of the three macro-areas, we use a distance-based test for homogeneity of multivariate dispersions.

Applying the method proposed in [2] to the three variance/covariance matrices estimated in model (2) and using the R package *vegan* (see [13]), we find that the means of the Euclidean distances between points and centroid within each group are 3.677 in the North, 4.238 in the Center and 6.329 in the South, showing that, as we saw below, the points of the South are more scattered. Similar results are obtained if we calculate the distances from the median within each group (respectively, 3.645, 4.217 and 6.303). Both the tests ANOVA (with centroids and medians) give p -values less than $2.2 * 10^{-16}$, proving that the three matrices are different, so that, there are different correlations between the school effects and different variance structures of random effects in the three macro-areas. If we repeat this study on the variance/covariance matrices of the errors we notice the same trend of the random effects' matrices: the distributions and the variances of the residuals of the three macro-areas are different, the distances between the points and the centroids within each group are about 14 in the North, 15 in the Center and 16 in the South. The test ANOVA gives a p -value less than $2.2 * 10^{-16}$. The high dispersion of the residuals in the South suggests that there is a considerable part of variability that remains unexplained.

3.2. Variables at school level across macro-areas

Now, it may be interesting to understand if the information at school level (number of students, percentage of female, immigrants..., private schools, etc.) are significant covariates in modeling the predicted $\hat{\mathbf{b}}_j$ of the random effects (school effects). We therefore proceed with a second step in the analysis, in which we regress the predicted school effects $\hat{\mathbf{b}}_j$ of model (2) against variables at school level, by means of three (bivariate) simple regression models, one for each macro-area. The variables at school level are divided into two groups: (i) the peers effects related to the composition of student body and (ii) managerial and

Table 5. ML estimates of model (12) fitted to data of Northern, Central and Southern area, with the only variables selected by the LASSO.

Model coefficients	North math	Center math	South math
Intercept	−1.467*	−3.899***	−3.946**
Mean ESCS			2.042***
Female percentage	0.038**	0.063**	0.066*
First generation imm perc	−0.025		
Second generation imm perc		0.146***	
Early-enrolled student perc			
Late-enrolled student perc	−0.057*		−0.208***
Number of classes			
Number of students	0.003**	0.004*	
Average num of stud per class			0.110*
Private school		−2.710**	
IC			
Model coefficients	North read	Center read	South read
Intercept	0.156	−0.380	−0.753
Mean ESCS	−1.727***	−0.396	1.095***
Female percentage			0.032*
First generation imm perc			
Second generation imm perc		0.153***	
Early-enrolled student perc		−0.203*	−0.097*
Late-enrolled student perc	0.025		−0.056*
Number of classes			
Number of students	0.002		
Average num of stud per class			
Private school	−1.424***	−2.457**	
IC			

Notes: Asterisks denote different levels of significance: $0.01 \leq p\text{-val} \leq 0.1$.

* $0.001 \leq p\text{-val} \leq 0.01$.

** $0.0001 \leq p\text{-val} \leq 0.001$.

*** $p\text{-val} \leq 0.0001$.

structural features of the school. The model is the following:

$$\hat{\mathbf{b}}_j^{(R)} = \boldsymbol{\gamma}_0^{(R)} + \sum_{h=1}^H \boldsymbol{\gamma}_h^{(R)} z_{jh}^{(R)} + \boldsymbol{\eta}_j^{(R)}, \quad (12)$$

where $R = \{\text{North}, \text{Center}, \text{South}\}$; $j = 1, \dots, J$ is the index of the school; $\hat{\mathbf{b}}_j^{(R)}$ is the predicted random effect of the j th school of models (2); $z_{jh}^{(R)}$ is the value of the h th predictor variable at school's level; $\boldsymbol{\gamma}^{(R)} = (\boldsymbol{\gamma}_0^{(R)}, \dots, \boldsymbol{\gamma}_H^{(R)})$ is the bivariate $(H+1)$ -dimensional vector of parameters; $\boldsymbol{\eta}_j^{(R)}$ is the zero mean gaussian error.

On this model we decide to perform a variable selection using the LASSO algorithm (see [17]). In fact, when the covariates are highly collinear this soft-threshold shrinkage method performs very well (see [10]). The penalization parameter λ that tunes the level of shrinkage and consequently the number of covariates to discard from the model is chosen by cross-validation techniques. Estimates of models (12) are reported in Table 5.

The composition of the school's peers, such as female, early/late-enrolled students percentage, weighs more in the South than in the North, in both the fields. The mean ESCS of the school is very significant and weighs positively in the South, while it weighs negatively in the North. Lastly, being a private school is significant just in the North and in the Center and it weighs negatively, suggesting that public schools are on average better than private ones.

Table 6. First line of the table shows the variance /covariance matrices of the two random effects and of the residuals estimated by model (13), while the second line reports the correlation coefficient between the two random (school or class) effects in reading and mathematics.

School	Class	Residuals
$\begin{pmatrix} 10.4 & 4.30 \\ 4.30 & 3.50 \end{pmatrix}$	$\begin{pmatrix} 17.4 & -1.02 \\ -1.02 & -18.4 \end{pmatrix}$	$\begin{pmatrix} 159.7 & 61.70 \\ 61.70 & 111.9 \end{pmatrix}$
cor = 0.712	cor = -0.05	

4. Bivariate three-level linear model: students nested in classes, nested in schools

At this point, we take into account the class level into the model, therefore fitting a three-level linear model in which pupils are nested in classes, that are in turn nested in schools. In that way, we can analyze how much of the random effects is really due only to the school and how much only to the class. Previous studies show that the main differences in educational attainments elapse within schools, and not between schools: attending certain classes weighs more than attending certain schools (see [11]). We fit bivariate three-level model in which pupil i , $i = 1, \dots, n_{lj}$; $n = \sum_{l,j} n_{lj}$ (first level) is in class l , $l = 1, \dots, L_j$; $L = \sum_j L_j$ (second level) that is in school j , $j = 1, \dots, J$ (third level):

$$y_{ilj} = \beta_0 + \sum_{k=1}^K \beta_k x_{kilj}^{(R)} + \mathbf{b}_j + \mathbf{u}_{lj} + \epsilon_{ilj}, \quad (13)$$

where y_{ilj} is the bivariate outcome with mathematics and reading achievements of pupil i , in class l , in school j ; $x_{kilj}^{(R)}$ is the value of the k th predictor variable at pupil's level; $\beta = (\beta_0, \dots, \beta_K)$ is the bivariate $(K+1)$ -dimensional vector of coefficients; $\mathbf{u} \sim N_2(\mathbf{0}, \Sigma_u)$ is the matrix of the two random effects (mathematics and reading) at class level; $\mathbf{b} \sim N_2(\mathbf{0}, \Sigma_b)$ is the matrix of the two random effects (mathematics and reading) at school level; $\epsilon \sim N_2(\mathbf{0}, W)$ is the error, with \mathbf{u} independent of ϵ , \mathbf{b} independent of ϵ , \mathbf{b} independent of \mathbf{u} , \mathbf{b}_m independent of $\mathbf{b}_n \forall m \neq n$ and \mathbf{u}_{pj} independent of $\mathbf{u}_{qj} \forall p \neq q$ and \forall school j .

Since the coefficients of the variables at student level are similar to the previous ones, we focus the attention on the random effects. Table 6 shows the variance/covariance matrices of the two random effects and of the residuals.

We notice that the variances of the class effect are higher than the school ones, both in mathematics and reading (17.4 vs. 10.4 in mathematics and 18.4 vs. 3.50 in reading), suggesting again that the effect of the class is stronger than the school one. While the class effects in the two fields have about the same variances, regarding the school the effect in mathematics is stronger than the reading one. Looking at the two correlation coefficients, it's clear that the effects of the school in the two fields are quite correlated (coef. 0.712) and they may represent better or worse schools, that give coherent contributes in the two topics. On the other hand, the two class' contributes are totally uncorrelated (coef. -0.05), so that there are classes that give a good contribute in reading and a bad one in mathematics and viceversa, probably depending on teachers.

Table 7. The table shows the variance/covariance matrices of the two random effects and of the residuals across macro-areas estimated by models (14) and the correlation coefficient of random (school or class) effects between reading and mathematics.

	North	Center	South
School	$\begin{pmatrix} 4.99 & 1.82 \\ 1.82 & 1.65 \end{pmatrix}$ cor = 0.63	$\begin{pmatrix} 5.89 & 3.49 \\ 3.49 & 3.56 \end{pmatrix}$ cor = 0.76	$\begin{pmatrix} 15.9 & 5.74 \\ 5.74 & 4.54 \end{pmatrix}$ cor = 0.67
Class	$\begin{pmatrix} 6.20 & -1.13 \\ -1.13 & 17.5 \end{pmatrix}$ cor = -0.10	$\begin{pmatrix} 13.9 & -0.88 \\ -0.88 & 18.1 \end{pmatrix}$ cor = -0.05	$\begin{pmatrix} 40.9 & 0.07 \\ 0.07 & 20 \end{pmatrix}$ cor = -0.002
Residuals	$\begin{pmatrix} 145.4 & 47.17 \\ 47.17 & 95.04 \end{pmatrix}$	$\begin{pmatrix} 164.6 & 63.14 \\ 63.14 & 114.8 \end{pmatrix}$	$\begin{pmatrix} 167.5 & 79.09 \\ 79.09 & 136.0 \end{pmatrix}$

Again, following the same reasoning about the geographical areas introduced in Section 4.1, we fit model (13) in each one of the three macro-areas:

$$\mathbf{y}_{ilj}^{(R)} = \boldsymbol{\beta}_0^{(R)} + \sum_{k=1}^K \boldsymbol{\beta}_k^{(R)} x_{kilj} + \mathbf{b}_j^{(R)} + \mathbf{u}_{lj}^{(R)} + \boldsymbol{\epsilon}_{ilj}^{(R)} \tag{14}$$

with $R = \{\textit{North}, \textit{Center}, \textit{South}\}$.

Table 7 shows the variance/covariance matrices of the two random effects and of the residuals, in the three areas.

Again, we see that the effects of the class are stronger than the school ones, indeed the variances of the class effects are higher. Regarding the school effects, the contribute in mathematics is always stronger than the reading one and overall the random effects weighs more in the South than in the North. Regarding the class effects, while in the North it seems that the class weighs more in reading than in mathematics, in the South it is the opposite. Again, the random effects in the South are stronger than in the North. Looking at the correlation coefficients we confirm that the two school effects are quite correlated in all the three macro-areas, while the two class effects are snugly uncorrelated.

Figure 3 shows the bivariate school and class effects predicted by model (13) and colored by macro-areas.

Regarding the school effects, it is clear that the variability of the value-added of schools in the South is much higher than the ones in the Center and in the North. This confirms that the impact the school has on students' performances in the South is stronger than the ones in the North and in the Center, that are similar. The same trend can be observed in the class effects, where the points of the South are more scattered, suggesting again that the effect of the class is stronger in the South than in the rest of Italy. Looking at the two shapes of points distributions, we see that there is a strong correlation (coefficient 0.712) between the school effects in reading and mathematics, that is most of the schools give coherent contributes in reading and mathematics, both positive or negative. On the other hand, there is no correlation (coefficient -0.05) between the two class effects. The class effect may be influenced by the teachers and a class may has a 'good' teacher of mathematics and a 'bad' one in reading or viceversa.

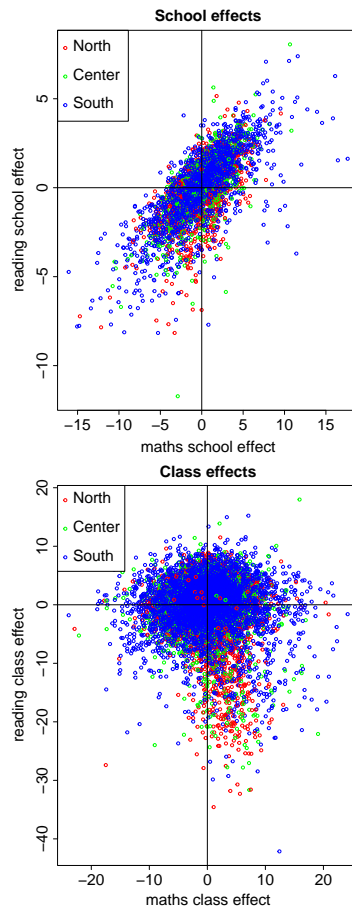


Figure 3. Plots of the school and class effects in mathematics and reading, predicted by model (13). Colors identify the three macro-areas: blue for the South, red for the North and green for the Center.

Now, it may be interesting to analyze if there is any kind of correlation between the school and class effects, such as if different kinds of schools (better or worse) contain different kinds of classes (better or worse). For this reason, from Figure 3 we can identify the best schools in the first quarter, where schools give positive values-added in both reading and mathematics, and the worst ones in the third quarter, where schools give negative values-added in both the topics. In the same way, we can identify better and worse classes. We have computed the mean percentage of virtuous classes, that is classes with the effects in the first quarter, in each sector of schools, such as schools with values-added in all the four quarters. Figure 4 shows the distribution of these percentages.

The school effect is snugly uncorrelated with the effects of the classes that the school contains, indeed there are good percentages of virtuous classes in all kinds of schools. This means that there are schools with a positive (or negative) value-added containing classes with negative (or positive) mean value-added, that is the goodness of a school is uncorrelated with the goodness of the contained classes.

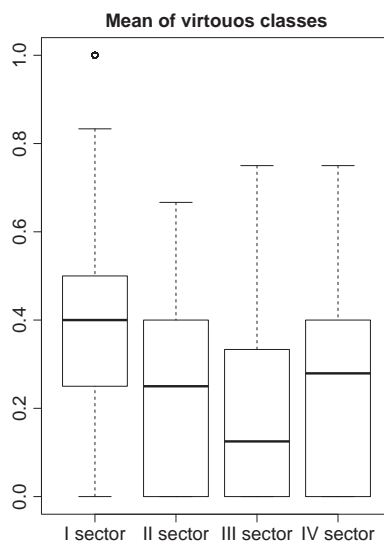


Figure 4. Boxplots of the mean percentages of virtuous classes stratified by sectors of schools.

5. Concluding remarks

This work explores how the students' achievements depend on students' characteristics and which may be the effects of attending specific schools. The data concerns INVALSI reading and mathematics test scores of students attending the first year of junior secondary school in the year 2012/2013.

The innovation of this work is in the use of bivariate multilevel models, that allow us to explore the interaction and the correlation between the effects of schools and classes on reading and mathematics achievements and to compare the associations between students' profile and their performances in the two fields. Previous studies have mainly analyzed school effects in reading and mathematics as separate, denying the possibility to have a complete view of the effectiveness of the school. With the bivariate models, we can predict school/class effects in both the topics, analyze how they are correlated, if the class effects are coherent within schools and if they depend on the same variables at school/class level.

Univariate multilevel models had been already applied in previous studies in order to explain how the reading or mathematics (separately) achievements depend on students' characteristics. From this point of view, the results obtained by the new approach are in line with existing research. What emerges are some recurrent associations between the outcome variable and regressors: females have, on average, better results than males in reading but worse one in mathematics, first and second generation immigrants have more difficulties than Italian students, especially in reading, and being early/late-enrolled student decreases the average result. Furthermore, the pupil's ESCS, index of the socioeconomic status of the student, has a strong positive influence on the achievement and the CS5 (the student's achievement at grade 5) is positively associated with the current score and this claims for a value-added (VA) specification of the statistical model. Lastly, from the models emerges that students of the Central and especially of the Southern Italy, have worse mean results than students of the North, showing big heterogeneities within the Country, that

brought to the need to have three different models. What is interesting is that students' characteristics are different across the three geographical macro-areas, Northern, Central and Southern Italy, which can be considered as three different educational systems. The variables at student level that more influence the CRS and CMS are heterogeneous across macro-areas: the ESCS is much more relevant in the South than in the North and being first and second generation immigrants decreases the mean result less in the South than in the North.

As anticipated, the real new contribution of this approach is in the estimates of the bivariate random effects, with their correlations. The school effect, defined as the effect of attending a specific school on a student's test score, is modeled as a random effect \hat{b}_j and is regressed against school level variables with the aim of characterizing the features of those schools that exert a positive/negative effect on academic performance. Even the school effects are different across macro-areas: in the South, they are much more scattered, suggesting that the school effect is much stronger. Therefore, while being private or public school influences the school effect in the North, in the South the mean ESCS of the school is one of the most relevant variables that adds positive value-added, showing that in the South the differences between schools tend to increase the inequalities between disadvantaged and advantaged students. Furthermore, it emerges that, in Italy, the school effect in mathematics is much stronger than the one in reading. In the macro-areas, this behavior occurs also in the South of Italy, while in the North is less pronounced. The correlation coefficients show a coherency between the school effects in the two topics, proving that generally the contributes of the school in reading and mathematics are positively correlated and this defines which are 'good' ('bad') schools. Therefore, it is possible to identify 'good' ('bad') schools, knowing that they give positive (negative) value-added in both the topics.

In the same way, also the class effects \hat{u}_{ij} are accounted for in the model, showing a trend that is similar to the one of school effects, even stronger. The main difference between the two effects is that the reading and mathematics class effects are not correlated like the school ones, denying the possibility to identify 'good' ('bad') classes. This arises from the fact that such kind of contributes at class level are probably given by the teachers, and students in a class may have a good teacher of mathematics and a less good one of reading or viceversa, without any kind of correlation. Anyway, the class effects follow similar trends of the school effects: the contributes in mathematics are more pronounced than in reading and in the South they are again stronger than in the North in both the topics, being different across macro-areas. At last, we see that it is impossible also to find a correlation between school and contained classes effects, that prove to be independent. This means that good (or bad) schools may contain bad (or good) classes. From the univariate cases, instead, we obtained that there was a dependence between school and classes (i.e. school with a positive value-added in mathematics/reading contained classes with a mean positive value-added in mathematics/reading). In the bivariate case, we loose this dependence because within each school, the contributes that classes give in mathematics and reading are very heterogeneous and not coherent with the contribute of the school. Therefore, the effectiveness of schools is independent from the effectiveness of the contained classes. We can therefore conclude that sometimes it is possible to identify and choose a good school, but within it there is still variability between and within classes and this variability changes across the three geographical macro-areas.

Further studies may be done to explore other aspects of the Italian educational system. It can be interesting to deepen the geographical differences, analyzing the districts; to explore if there is a sort of homogeneity of the variables within the schools and within the classes; to discover how much the teachers influence the class effects; to provide a way to treat the missing data and, particularly, to explore if there is a way to reduce the geographical heterogeneity, in order to provide a good educational level for all Italian students.

5.1. Policy and managerial implications

The results of this analysis bring to the need to make some considerations about policy and managerial implications.

First of all, the use of school-effect estimates should be carried out in a proper way. As we have pointed out before, the contributes that schools provide in both reading and mathematics may be different and, sometimes, even opposite. It can be the case that the school-effect is positive in mathematics and negative in reading, or viceversa. In this sense, policy makers should be clear about what they want to promote and reward. Sanctions and rewards can induce substantial consequences. For this reason, we should be sure about how the estimates are robust and stable, so that reliable to imply such policy and managerial consequences (see [19]).

Secondarily, more research is needed for brightening the factors behind the differences in the schools' profiles and for explaining the coherence/incoherence between the effects in mathematics and reading within schools. As we have seen, the two effects are affected by different variables. This heterogeneity may be due to various aspects, such as students' attitudes, teachers' effectiveness or schools' strategies and activities.

Lastly, our findings highlight that schools are intrinsically multi-output organizations. Our study is a first step in this direction and in the next future we will concentrate on typical implications for multi-output organizations, such as checking for economies of scope and developing new indicators for unmeasured outputs (as non-cognitive skills). Given these three evidences, the policy implication consists in the implementation of an evaluation system that considers the role of the school in affecting achievement in the tested disciplines reading and mathematics. Until now, there is not a relationship between the results of INVALSI tests and consequences for the schools, such as funding allocations or reputational information as league tables, etc. In our opinion, statistically predicted school effects should be gradually introduced in the educational system, to stimulate schools in pursuing better performances. Of course, the actors should be aware of the methodological cautions expressed above, especially about the variance in school effects between the different disciplines, and about how the effect itself is different in North vs. South Italy. Another policy implication is that predictions of class level effects should be experimented in a more systematic way, and information about teachers and classes should be improved to better understand whether the class effect is related to compositional differences or teachers' quality in terms of actions from policy-makers, different results would lead to alternative areas of interventions. Finally, INVALSI should promote tests also in additional disciplines, as for instance science and English literacy, with the aim of verifying how heterogeneous is the effect of schooling on a broader range of knowledge areas.

Acknowledgments

The authors are grateful to INVALSI for having provided the original data set.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work is conducted within FARB – Public Management Research: Health and Education Systems Assessment, funded by Politecnico di Milano.

ORCID

F. Ieva  <http://orcid.org/0000-0003-0165-1983>

References

- [1] T. Agasisti, F. Ieva, and A.M. Paganoni, Heterogeneity, school-effects and achievement gaps across italian regions: Further evidence from statistical modeling, MOX, Dipartimento di Matematica F. Brioschi, Politecnico di Milano. Available at <https://www.mate.polimi.it/biblioteca/add/qmox/07-2014.pdf> (31 January 2014).
- [2] M.J. Anderson, *Distance-based tests for homogeneity of multivariate dispersions*, Biometrics 62 (2006), pp. 245–253.
- [3] K.E. Castellano, S. Rabe-Hesketh, and A. Skrondal, *Composition, context, and endogeneity in school and teacher comparisons*, J. Educ. Behav. Stat. 39 (2014), pp. 333–367.
- [4] J. Dronkers and P. Robert, *Differences in scholastic achievement of public, private government-dependent, and private independent schools a cross-national analysis*, Educ.Policy 22 (2008), pp. 541–577.
- [5] J. Fox, *Linear mixed models*, Appendix to An R and S-PLUS Companion to Applied Regression, 2002. Available at <http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-mixed-models.pdf>.
- [6] J. Fox and S. Weisberg, *An R Companion to Applied Regression*, Sage, London, 2010.
- [7] A.R. Gilmour, B.R. Cullis, S.J. Welha, and R. Thompson, *2002 asreml Reference Manual 2nd Edition*, release 1.0 nsw. Agriculture Biometrical Bulletin 3, NSW Agriculture, Locked Bag, Orange, NSW 2800, Australia.
- [8] E.A. Hanushek, *Education production functions*, in *International Encyclopaedia of Economics of Education*, M. Carnoy, ed., Pergamon Press, Tarrytown, NY, 1995, pp. 277–282.
- [9] E.A. Hanushek and L. Woessmann, *The economics of international differences in educational achievement*, in *Handbook of the Economics of Education*, Vol. 3, E. A. Hanushek, S. Machin, and L. Woessmann, eds., Elsevier B.V., Amsterdam, 2011, pp. 91–192.
- [10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, New York, 2009.
- [11] C. Masci, *Bivariate multilevel models for the analysis of reading and maths pupils' achievements*, Tesi di laurea magistrale, Politecnico di Milano, 2013/2014.
- [12] C. Masci, F. Ieva, T. Agasisti, and A.M. Paganoni, *Does class matter more than school? Evidence from a multilevel statistical analysis on Italian junior secondary school students*, Socio-Econ. Plann. Sci. 54 (2016), pp. 47–57.
- [13] J. Oksanen, F.G. Blanchet, R. Kindt, P. Legendre, P.R. Minchin, R.B. O'Hara, G.L. Simpson, P. Solymos, M.H.H. Stevens, and H. Wagner, *vegan: Community Ecology Package*, 2013. R package version 2.0-10.
- [14] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Core Team, *nlme: Linear and Nonlinear Mixed Effects Models*, 2014, R package version 3.1-117.

- [15] J.J.C. Pinheiro and D.M. Bates, *Mixed-effects Models in S and S-PLUS*, Springer, New York, 2000.
- [16] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014.
- [17] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Ser. B Methodol. 58 (1996), pp. 267–288.
- [18] A. Vignoles, R. Levacic, J. Walker, S. Machin, and D. Reynolds, *The Relationship between Resource Allocation and Pupil Attainment: A Review*, Centre for the Economics of Education, London School of Economics and Political Science, 2000.
- [19] J.D. Willms and S.W. Raudenbush, *A longitudinal hierarchical linear model for estimating school effects and their stability*, J. Educ. Meas. 26 (1989), pp. 209–232.