JOURNAL OF APPLIED QUANTITATIVE METHODS

The International Conference
"Innovation and Society 2011. Statistical Methods
for the Evaluation of Services (IES 2011)"

DIFFERENTIAL VARIABILITY OF TEST SCORES AMONG SCHOOLS: A MULTILEVEL ANALYSIS OF THE FIFTH-GRADE INVALSI TEST USING HETEROSCEDASTIC RANDOM EFFECTS

Claudia SANI [1]

MSc in Statistical Science,
Department of Statistics,
University of Florence

Data Analyst,
IMS Health, Milan, Italy

E-mail: claudia.sani86@gmail.com

Leonardo GRILLI [2]

Associate Professor of Statistics,
Department of Statistics "*G. Parenti*",
University of Florence

E-mail: grilli@ds.unifi.it

Abstract: *The performance of a school system can be evaluated through the learning levels of the pupils, usually summarized by school mean scores. The variability of the mean scores among schools is rarely studied in detail, though it is a crucial issue especially in primary schools: in fact, a high variability among schools raises doubts on the capacity of the system to guarantee equal educational opportunities. To investigate the patterns of variability in Italy, we analyse data from INVALSI, the Italian national institute for the evaluation of the school system, which regularly carries out standardized tests to assess the learning levels of the pupils at various grades. We consider the mathematics test administered to fifth-grade pupils at the end of the 2008/2009 year, along with a pupil's questionnaire for measuring socio-economic factors. The analysis is performed using a random intercept linear model on the Rasch score of the mathematics test, with pupil-level errors depending on gender and school-level errors depending on the geographical area. The model includes several demographic and socio-economic explanatory variables and some compositional variables obtained as school means of pupil variables. The results show a considerable increase in the residual variance among schools when going from North to South, pointing out a serious issue of fairness in Southern Italy. The situation is mitigated by the finding that a substantial part of the residual variance among schools is due to a few schools with exceptionally positive results.*

Key words: *Achievement, Multilevel model, Outlier, Rasch score, School performance*

1. Introduction

The evaluation of the performance of the educational system using statistical methods is becoming increasingly important. In Italy, the evaluation is in charge of the Italian national institute for the evaluation of the school system (INVALSI), whose primary purpose is to gather data and provide tools for the evaluation of the school system as a whole. These data would enable policy makers, administrators and citizens to establish if the Italian school system is achieving its objectives.

One of the main goals of a public education system is to offer equal educational opportunities to all students. This equity, especially in primary school, should be ensured by a low variability among schools. Such a feature is crucial, though it is rarely studied in detail.

In this paper we wish to identify and quantify the determinants of pupil achievement in Italian primary schools, distinguishing between individual factors (demographic, social, economic and cultural) and contextual factors (compositional variables defined as school means of pupil variables). To this end, we adopt a multilevel regression model, which properly accounts for the hierarchical structure of data (pupils nested into schools) by partitioning the residual variance into pupil and school components. We use a random intercept linear model with heteroscedastic variance components: in particular, the school residual variance is allowed to change with the geographical area. This feature is crucial in order to investigate the issue of differential variability of achievement across schools. In fact, a descriptive analysis shows that the schools in Northern regions are similar in terms of mean achievement score, whereas the schools in Southern regions are very heterogeneous, with excellent schools beside bad schools: this is a key fact which needs to be investigated through a statistical model controlling for the available explanatory variables.

Our analysis refers to the Rasch score of the math test administered by INVALSI in May 2009 to a sample of pupils attending the fifth grade of primary school. The sample consists of approximately 1000 schools and 40000 students. The same data were analysed by Petracco-Giudici, Vidoni and Rosati (2010) through a multilevel random slope model with homogeneous variances in a study of compositional effects.

This paper summarizes and develops the dissertation of Sani (2011). The rest of the article is organized as follows: in Section 2 we describe the data and summarize the preliminary analysis; in Section 3 we outline the multilevel regression model and report the estimation results; finally, in Section 4 we discuss the main findings.

2. Data and preliminary analysis

We considered the survey conducted by Italian national institute for the evaluation of the school system (INVALSI) carried out at the end of the year 2008-09. The aim is to evaluate the proficiency in Italian and mathematics of pupils attending the second grade and the fifth grade. The test was administered by external staff to the pupils of a probabilistic sample of schools. Here we focus on the math test for fifth grade pupils (about 11 years old). The test includes 41 multiple-choice items, summarized by the Rasch score (Andrich, 1988). The pupils also filled a questionnaire for collecting variables which are proxies of the social, economic and cultural conditions of their families. Other data were collected through the school secretaries. A report on the survey is available on the web page of the institute (INVALSI 2009).

After a step of data cleaning (details in Sani 2011), the sample for our analysis includes 38708 pupils belonging to 932 schools. The number of pupils per school varies from 1 to 123 with a median of 37.

The analysis was based on the following variables (mostly taken from the pupil questionnaire):

- *Demographic variables*: gender, foreigner, year of birth;
- *Socio-cultural variables*: people with whom the pupil lives, language spoken at home, availability of a computer or an encyclopaedia or an internet connection, number of books at home, help with homework, hours playing video games, hours reading;
- *Wealth*: presence of alarm at home, number of bathrooms at home, number of cars at home;
- *School climate*: unease score (integers from 0 to 4);
- *Geographical area of the school*: North-West (Valle d'Aosta, Piemonte, Liguria, Lombardia), North-East (Trento, Bolzano, Veneto, Friuli Venezia Giulia, Emilia Romagna), Center (Toscana, Umbria, Marche, Lazio), South (Abruzzo, Molise, Campania, Puglia), South-Isles (Basilicata, Calabria, Sicilia, Sardegna). The South-Isles area is consistent with the definition adopted by major international surveys such as PISA and TIMSS.

Some key statistics about the pupils are as follows: 50.8% are males; 94.1% were born in Italy; about 90% were born in 1998 (regular students); as for the geographical area, 23.4% attend a school in North-West, 22.7% in North-East, 21.2% in Center, 16.4% in South and 16.3% in South-Isles.

The Rasch score on the math test lies between -5.664 and 4.683, with mean 0.009 and standard deviation 0.944; the quartiles are -0.626, -0.135, and 0.515. The distribution of the math score in the five geographical areas is similar, except for the substantially lower mean value in South-Isles (-0.270 with respect to Center). However, the differences among areas become relevant if one aggregates the score by school and looks at the distribution of the school mean score (Figure 1): it appears that the schools in the South and South-Isles are notably more heterogeneous, with several positive outliers (the standard deviations are North-West 0.30, North-East 0.34, Center 0.41, South 0.67, and South-Isles 0.54).

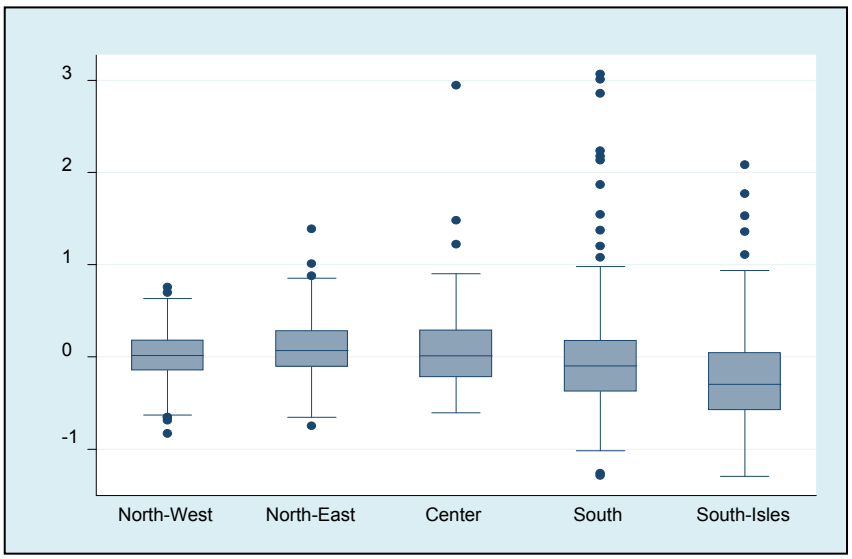

Figure 1. Distribution of school mean Rasch math score by geographical area

A preliminary evaluation of the effects of the explanatory variables on the math score has been carried out using the analysis of variance with Scheffé tests for multiple comparisons. Significantly higher mean scores have been detected for: males, Italian pupils, pupils speaking Italian at home, pupils helped by the family or not needing help, pupils reading at home, pupils playing video games, pupils with the availability of a computer or an encyclopaedia or an internet connection; other variables with positive effect are the number of books at home, the number of cars, the number of bathrooms, the presence of an alarm. Finally, the unease score (measuring the perceived unease of the pupil at school) has a negative effect.

Table 1 reports the definitions of the explanatory variables considered in the multilevel models. They were chosen on the basis of the preliminary analysis, which also suggested how to merge the categories of some variables. All the variables in Table 1 refer to the pupils since the dataset does not include any school feature.

Table 1. Description of the pupil variables

Variable	Description
DEMOGRAPHIC	
Female	1: female; 0: male
Foreigner	1: born abroad; 0: born in Italy
Year of birth1	1: born before 1998 (behind schedule); 0: otherwise
Year of birth 2	1: born in 1998 (regular); 0: otherwise
Year of birth 3	1: born after 1998 (in advance); 0: otherwise
Family	1: he/she lives with a single parent or with other persons; 0: he/she lives with both parents
SOCIO-CULTURAL	
Computer	1: he/she has a computer at home; 0: otherwise
Encyclopaedia	1: he/she has an encyclopaedia at home; 0: otherwise
Internet	1: he/she has Internet at home; 0: otherwise
Reading	1: he/she reads books or comics as pastime; 0: otherwise
Help with homework	1: he/she is helped with homework by family members, or he/she declares that no help is needed; 0: otherwise
Videogames	1: he/she does not play videogames; 2: he/she plays less than 1 hour; 3: he/she plays 1-2 hours; 4: he/she plays more than 2 hours
Books+100	Number of books at home /100 (this variable is obtained by replacing the classes of the questionnaire 0-10, 11-25, 26-100, 101-200, 200+ with their midpoints 5, 18, 63, 150, 300; the values are then divided by 100 in order to interpret the regression coefficient as the effect of +100 books)
WEALTH	
Alarm	1: he/she has an alarm at home; 0: otherwise
Bathrooms2+	1: he/she has two or more bathrooms at home; 0: otherwise
Cars	Number of cars (3 stands for 3 or more)
SCHOOL CLIMATE	
Uneasiness	Number of affirmative responses to the questions q13a, q13b, q13c, q13d (range 0 to 4)

3. Multilevel analysis

A suitable regression model for the analysis of the Rasch score of the math test is a random intercept two-level linear model with pupils at level 1 and schools at level 2 (Goldstein, 2010; Snijders and Bosker, 2011). This model allows us to study the effect of both individual-level and school-level explanatory variables, while accounting for the

correlation of the responses of the pupils of the same school. We use a linear model with normally distributed heteroscedastic errors at both levels. The issue of heteroscedasticity, which is well-known in principle but often neglected in applied research, is fundamental for our purpose. In our case the heteroscedasticity is due to categorical variables such as gender and geographical area, thus the model can be written as:

$$Y_{ij} = \alpha + \sum_{r=1}^{R} \beta_r X_{rij} + \sum_{s=1}^{S} \gamma_s Z_{sj} + u_j^{(k)} + e_{ij}^{(m)}$$

where Y_{ij} is the response variable for the *i*-th level 1 unit (pupil) of the *j*-th level 2 unit (school), X_{rij} are level 1 explanatory variables, Z_{sj} are level 2 explanatory variables, $e_{ij}^{(m)}$ are level 1 errors (with m=1,...,M strata of level 1 units), and $u_j^{(k)}$ are level 2 errors or random effects (with k=1,...,K strata of level 2 units). The model errors are assumed to be normally distributed with zero mean, whereas their variance is different across strata; in the application, the strata of level 1 units (pupils) are defined by gender (M=2), while the strata of level 2 units (schools) are defined by the geographical area (K=5).

In order to study the school-level heteroscedasticity, the schools could be stratified according to smaller areas such as the regions. However, we used the five macro areas because they yield a balanced structure with about 180 schools per area (except for the higher number in North-West, see Table 5 later): this ensures that the five area-specific variances are estimated with good precision, and such precision is nearly constant across areas.

The role of the hierarchy can be summarized by the Intraclass Correlation Coefficient (ICC), which is defined as the ratio of the level 2 variance on the total variance (level 2 plus level 1). In our application, the model has heteroscedastic errors at both levels, thus there are $M \times K$ versions of the ICC: denoting with $\sigma^2_{e^{(m)}}$ the variance of the level 1 errors in stratum m and $\sigma^2_{u^{(k)}}$ the variance of the level 2 errors in stratum k, the ICC for strata m and k is

$$ICC^{(m,k)} = \frac{\sigma^2_{u^{(k)}}}{\sigma^2_{u^{(k)}} + \sigma^2_{e^{(m)}}}$$

The model was fitted with Maximum Likelihood using the 'xtmixed' command of Stata (Stata Corp. 2009; Rabe-Hesketh and Skrondal 2008). The statistical significance at 5% was established using the Wald test for the regression parameters and the Likelihood Ratio test for the variances.

The level 1 variables considered in the models, selected on the basis of the preliminary analysis, were listed in Table 1. The dataset does not include any school feature such as public/private or urban/rural. The level 2 variables are the geographical area (coded with four dummy variables) and the compositional variables defined as school means of the pupil variables (labelled *SM-Female*, *SM-Uneasiness*, etc.); the compositional variables, which play a central role in multilevel theory, account for the environment of the school.

The model selection procedure consisted of three steps: models without explanatory variables, inclusion of level 1variables, and inclusion of level 2 variables. The variables were tried one at a time with a forward strategy guided by theory and preliminary results.

3.1. Multilevel models without explanatory variables

We first considered models without explanatory variables. The model with homoscedastic errors yields an estimated ICC equal to 21.1%: overall, 21.1% of the variance of the math score is due to the schools.

The next step concerned the heteroscedasticity of the errors. We found evidence of level 1 heteroscedasticity depending on gender (estimated standard deviations: males 0.894, females 0.811) and level 2 heteroscedasticity depending on geographical area (estimated standard deviations: North-West 0.241, North-East 0.259, Center 0.363, South 0.665, South-Isles 0.521). In terms of ICC the differences are astonishing: for example, for females the ICC is 8.1% in North-West and 40.2% in South (note this model has 2×5 versions of the ICC). Therefore, in the South a large part of the variability of the math score is due to the schools, thus there are bad schools beside excellent schools.

3.2. Multilevel models with explanatory variables

The model selection went on by adding level 1 explanatory variables. In the end, the selected variables are *Female*, *Foreigner*, *Encyclopaedia*, *Reading*, *Help with homework*, *Books+100*, *Bathrooms2+*, and *Uneasiness*.

Finally, the level 2 explanatory variables were inserted. The significant variables are the dummies for the geographical area and four compositional variables (school means): *SM-Encyclopaedia*, *SM-Reading*, *SM-Bathrooms2+*, *SM-Uneasiness*. The estimates for the final model are reported in Table 2.

Table 2. Multilevel linear model with explanatory variables at pupil and school level and heteroscedastic errors – Results of Maximum Likelihood estimation

Coefficient	Estimate	Std. Err.	Z	P>|Z|	[95% Conf. Interval]	
Constant	-0.589	0.142	-4.15	0.000	-0.866	-0.311
DEMOGRAPHIC						
Female	-0.113	0.009	-12.690	0.000	-0.130	-0.095
Foreigner	-0.324	0.019	-16.930	0.000	-0.361	-0.286
SOCIO-CULTURAL						
Encyclopaedia	0.039	0.010	3.800	0.000	0.019	0.060
Reading	0.039	0.011	3.430	0.001	0.017	0.061
Help with homework	0.047	0.015	3.220	0.001	0.018	0.076
Books+100	0.028	0.005	5.550	0.000	0.018	0.038
WEALTH						
Bathrooms2+	0.024	0.01	2.470	0.013	0.005	0.043
SCHOOL CLIMATE						
Uneasiness	-0.019	0.006	-3.180	0.001	-0.031	-0.007
GEOGRAPHICAL AREAS						
North-West	-0.004	0.033	-0.110	0.915	-0.068	0.061
North-East	0.062	0.035	1.760	0.078	-0.007	0.131
South	-0.009	0.058	-0.160	0.877	-0.124	0.106
South-Isles	-0.246	0.048	-5.150	0.000	-0.340	-0.153
COMPOSITIONAL VARIABLES (SCHOOL MEANS)						
SM-Encyclopaedia	0.509	0.127	4.000	0.000	0.260	0.759
SM-Reading	0.323	0.132	2.440	0.015	0.064	0.583
SM-Bathrooms2+	0.250	0.072	3.470	0.001	0.109	0.392
SM-Uneasiness	-0.433	0.067	-6.430	0.000	-0.564	-0.301
STANDARD DEVIATIONS OF LEVEL 2 ERRORS (RANDOM EFFECTS)						
Sd(North-West)	0.197	0.014			0.170	0.227
Sd(North-East)	0.237	0.017			0.205	0.274
Sd(Center)	0.351	0.023			0.309	0.399
Sd(South)	0.652	0.037			0.583	0.729
Sd(South-Isles)	0.486	0.029			0.432	0.545

STANDARD DEVIATIONS OF LEVEL 1 ERRORS				
Sd(Males)	0.886	0.005	0.877	0.895
Sd(Females)	0.805	0.004	0.796	0.813

The demographic variables have the strongest effects: holding other variables constant, females have a mean score 0.113 points lower than males, whereas foreigners have a mean score 0.324 points lower than Italians.

As for the socio-cultural variables, we find positive effects of *Encyclopaedia*, *Reading*, *Help with homework*, *Books+100*, while the effect of having a computer or an internet connection is not significant. The variable *Bathrooms2+*, which is a proxy of the family wealth, is highly significant. These variables reassert the key role of the family for the pupil achievement.

As for the school climate, the variable *Uneasiness* has a negative coefficient −0.019; its magnitude is moderate since the variable has a mode in 0 and it rarely takes values larger than 2.

The estimates for the dummies of the geographical areas confirm the low performance of South-Isles (−0.246 points with respect to Center).

The school means are intended to capture the compositional effects. The school mean of a binary variable is a proportion, thus +1 corresponds to +100% and +0.1 corresponds to +10%. For example, let us consider two hypothetical pupils with identical explanatory variables and errors, except that they attend two schools with a difference of 10% in the percentage of pupils having an encyclopaedia at home: it follows that the difference in the expected math score of the two pupils is the coefficient of *SM-Encyclopaedia* multiplied by 0.1, namely $0.509 \times 0.1 = 0.0509$ points. This is similar to the effect of individual socio-cultural variables. The three school proportions (*SM-Encyclopaedia*, *SM-Reading*, *SM-Bathrooms2+*) have positive coefficients: as expected, a good environment is beneficial for the achievement of any pupil. The fourth compositional variable (*SM-Uneasiness*) is the school mean of the uneasiness score, which has, as expected, a negative effect (−0.433). To appreciate the role of this variable, note that it has a distribution ranging from 0 to 1.238 with mean 0.476 and standard deviation 0.175: therefore, a difference of one standard deviation is associated to a difference of $-0.433 \times 0.175 = -0.0758$ points.

The level 1 errors have estimated standard deviations higher than the level 2 errors. The 2×5 estimated ICCs are reported in Table 3: even controlling for the explanatory variables, the proportion of residual variance due to the schools is still remarkably different across the areas. For example, for females the ICC is 5.7% in North-West, 39.6% in South and 26.7% in South-Isles. The values for South and South-Isles are extremely high: therefore, in those areas the pupil achievement is strongly affected by unobserved school-level factors.

Table 3. Estimated ICCs (variances in parenthesis)

	North-West (0.039)	**North-East (0.056)**	**Center (0.123)**	**South (0.425)**	**South-Isles (0.236)**
Males (0.785)	4.7%	6.7%	13.5%	35.1%	23.1%
Females (0.648)	5.7%	8.0%	16.0%	39.6%	26.7%

3.3. School rankings

A school random effect can be interpreted as the effectiveness of the school adjusted for the available explanatory variables. It is interesting to compare the (raw) ranking of the schools based on the mean math score with the (adjusted) ranking based on

the Empirical Bayes predictions of the random effects. The two rankings are very similar, especially in the top, thus the explanatory variables contribute little to account for the differences in the school performances. Both best schools and worst schools belong to South and South-Isles areas. This is a consequence of the higher variability. The good results are likely to be genuine since the test was administered by external staff; moreover, it is reassuring to note that the top schools have a standard deviation of the math score close to the value in the whole sample (0.944).

3.4. Importance of the types of determinants

The multilevel model assumes that the math score of a pupil is determined by observed pupil variables, observed school variables, unobserved pupil factors (level 1 error) and unobserved school factors (level 2 error). To understand the importance of each type of determinant, we computed the expected math score for some hypothetical profiles. As for the *observed variables*, the profiles are defined as follows:

- *Privileged pupil*: male, Italian, with encyclopaedia, hobby for reading, about 300 books, two or more bathrooms, he is helped with homework by family members or he does not need help, unease score equal to 0;
- *Underprivileged pupil*: female, foreigner, hobby for reading, about 5 books and at most one bathroom, she is not helped with homework, unease score equal to 2;
- *Effective school*: located in North-East, it is at the 95th percentile for the proportion of pupils holding an encyclopaedia, for the proportion of pupils with hobby for reading and for the proportion of pupils with two or more bathrooms, whereas it is at the 5th percentile for the mean unease score;
- *Ineffective school*: located in South-Isles, it is at the 5th percentile for the proportion of pupils holding an encyclopaedia, for the proportion of pupils with hobby for reading and for the proportion of pupils with two or more bathrooms, whereas it is at the 95th percentile for the mean unease score;

As for the *unobserved factors*, the profiles of the students are based on the distribution of the level 1 residuals, whereas the profiles of the schools are based on the distribution of the level 2 residuals (Empirical Bayes predictions of the random effects). Since the distributions of the model errors are assumed to be normal with zero mean, the 5th and 95th percentiles are at ±1.96 standard deviations. For example, an ineffective school in area *k* has a level 2 error equal to $-1.96\hat{\sigma}_u^{(k)}$.

Table 4 presents the variation in the expected math score when changing profile: for example, a privileged pupil has an expected score 0.707 points higher than an underprivileged pupil as for the observed variables, while the advantage is 3.544 points as for the unobserved factors. Note that the impact of the unobserved factors at the school level depends on the geographical area since the standard deviation of the random effects changes with the area.

Table 4. Expected math score for different profiles of pupil and school

Area	OBSERVED VARIABLES		UNOBSERVED FACTORS	
	Pupil underprivileged → privileged	School ineffective → effective	Pupil underprivileged → privileged	School ineffective → effective
North-West	+0.707	+0.569	+3.544	+0.788
North-East	+0.707	+0.569	+3.544	+0.948
Center	+0.707	+0.569	+3.544	+1.404
South	+0.707	+0.569	+3.544	+2.608
South-Isles	+0.707	+0.569	+3.544	+1.944

The figures of Table 4 show that the unobserved factors have a preeminent role, especially at the pupil level: as usual in educational research, the observed variables explain a minor part of the variation in the achievement score. As for the schools, in our application the observed variables are just the dummies for the areas and a few compositional variables, without any school feature such as public/private or urban/rural: in the light of this lack, the observed variables have a surprisingly relevant effect. The unobserved school factors play a key role in Southern regions, where their effect is much greater than the effect of observed variables.

3.5. Outlying schools

In this application there are several outlying schools, namely schools with an extreme value of the predicted random effect, corresponding to an exceptionally bad or good performance. Specifically, we define a school to be outlying when its value lies outside the whiskers of the box-plot (Figure 2), which are 1.5 times the interquartile range away from the first and third quartiles. Almost all outliers are positive and concentrated in South and South-Isles (Table 5). The South area has the highest number of outliers (14, corresponding to 8% of the schools) despite having the largest standard deviation of the random effects. Indeed, the histograms of Figure 3 show that South and South-Isles have markedly asymmetric distributions, raising doubts on the normality assumption of the random effects.

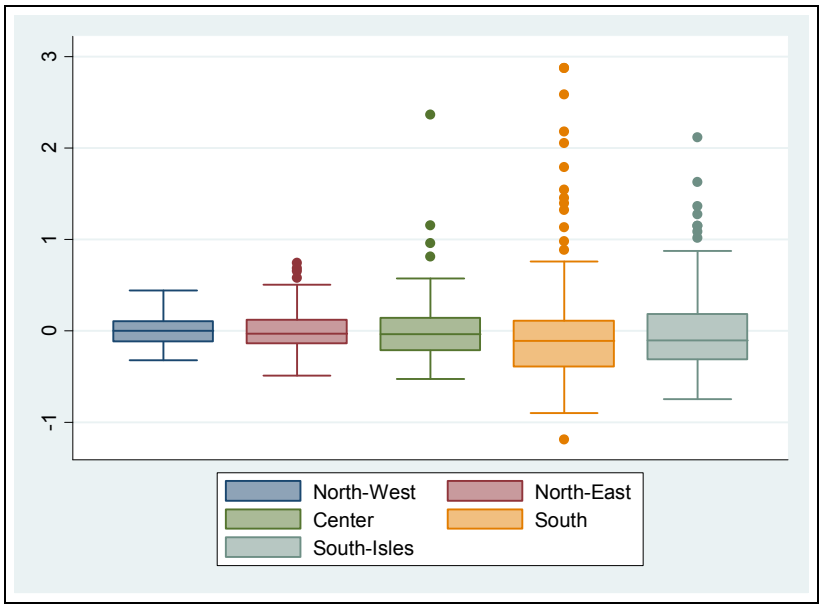

Figure 2. Box-plots of predicted random effects by geographical area (before eliminating outlying schools)

Table 5. Outlying schools by geographical area

Area	Sample	Positive outliers	Negative outliers	Total outliers	Percentage
North-West	216	0	0	0	0.0%
North-East	179	5	0	5	2.8%
Center	186	4	0	4	2.2%
South	175	13	1	14	8.0%
South-Isles	176	8	0	8	4.5%
Total	932	30	1	31	3.3%

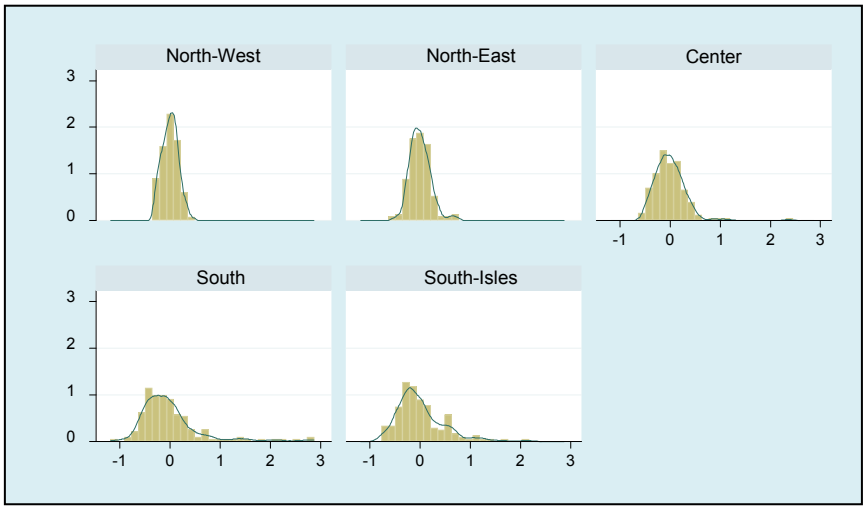

Figure 3. Histograms and kernel densities of predicted random effects by geographical area (before eliminating outlying schools)

To evaluate the role of the outlying schools, we removed those schools from the dataset and fitted the model again. Without outlying schools, the distributions of the predicted random effects are nearly symmetric in all areas. Moreover, the standard deviations of the random effects are substantially reduced in Southern regions (Table 6): in particular, in the South area the estimated standard deviation is halved and the ICC goes from an extraordinary 35.1% to a more common 11.9%. Therefore, a large part of the higher variability among the schools in Southern regions is due to a few outstanding schools.

Table 6. Standard deviation of random effects and ICC for males by geographical area: estimates with and without outlying schools

Area	Fraction of outlying schools	Model with *outliers*		Model without *outliers*	
		Std. Dev.	ICC	Std. Dev.	ICC
North-West	0/216	0.197	4.7%	0.197	4.7%
North-East	5/179	0.237	6.7%	0.179	4.0%
Center	4/186	0.351	13.5%	0.245	7.2%
South	14/175	0.652	35.1%	0.324	11.9%
South-Isles	8/176	0.486	23.1%	0.363	14.5%

4. Discussion

A public education system should offer equal educational opportunities to all students, especially in primary school. In Italy, the attainment of this target seems to be poor in Southern regions due to a very high variability among the mean performances of the schools. In this regard, the evidence from a descriptive analysis of the math score from the INVALSI test on fifth-grade pupils is confirmed by the results of a complex multilevel regression model with heteroscedastic variance components. The model adjusts the observed math score for both individual and compositional variables, while allowing the pupil-level variance to change with gender and the school-level variance to change with the geographical area. The proportion of residual variance due to the schools is remarkably different across the areas: it is 4.7% in North-West, 6.7% in North-East, 13.5% in Center, 35.1% in South and 23.1% in South-Isles (values for males). The high percentages of the Southern regions imply that in those regions the pupil achievement is strongly affected by unobserved school-level factors, insomuch that both the best and the worst schools in Italy are located in Southern regions.

However, a look at the Empirical Bayes predictions of the random effects reveals that in Southern regions there are several outlying schools, almost all with positive values. In particular, in the South area (Abruzzo, Molise, Campania, Puglia) there are 14 outlying schools (13 positive and 1 negative), corresponding to 8% of the schools of the area: when the model is fitted again without outlying schools, the proportion of residual variance due to the schools goes from 35.1% to 11.9%. In the South-Isles area (Basilicata, Calabria, Sicilia, Sardegna) the outlying schools (all positive) are 4.5% and their deletion entails a reduction of the residual variance from 23.1% to 14.5%. Therefore, a large part of the higher variability among the schools in Southern regions is due to a few schools with exceptionally positive results: this makes the overall picture less alarming than what might appear at first sight.

The estimates of the regression coefficients are in line with the expectations: on average, the math score is lower for foreigners, females and pupils feeling uneasy at school, whereas it is positively influenced by the economic, social and cultural background of the pupil. Moreover, there is evidence of substantial peer and contextual effects, which are partly captured by several significant compositional variables (school means of individual variables). Unfortunately, the available data do not include any school-level variable, such as the number of students per teacher or the presence of school facilities, thus hindering the analysis of contextual effects.

Our analysis could be extended to the test score on Italian language in order to establish similarities and differences with the findings on the math test. A joint analysis of the two test scores could be performed with a bivariate multilevel model (Goldstein 2010). It would be also interesting to analyse the test scores for several years to see whether the results are stable over time.

As for the methodological facet, the multilevel model we used for the analysis could be refined in two directions. The first one entails extending the model to more than two levels, for example using the province as the third level; this would require having access to data on the locations of the schools (not available in the dataset released by INVALSI). The other refinement entails using random effects with asymmetric distributions (Liu and Dey 2008) to account for the schools with exceptionally positive results.

In order to plan effective interventions to improve the school system, the issues highlighted by the statistical analysis should be investigated by experts in education. For

example, what are the reasons for the high variability of the performance of the schools in Southern Italy? A possible explanation relies on self-selection processes of pupils and teachers driven by unobserved variables. Since a large part of this variability is due to a few outstanding schools, the self-selection processes seem to be asymmetric ('positive' selection into excellent schools more than 'negative' selection into worse schools). As for the pupils, it is likely that educated and/or motivated parents are able to choose the best school in the neighbourhood. Therefore, we conjecture that the role of the family is not appropriately accounted for by the available explanatory variables. As for the teachers, in Southern regions there is a considerable mobility, with highly motivated teachers escaping from schools in difficult environments and moving toward good schools: such behaviour bolsters the differences among schools. Some work in the field is needed to understand the functioning of the mentioned self-selection processes of students and teachers. The inquiry should begin by a close look at both excellent and weak schools.

References

1. Andrich D. **Rasch Models for Measurement**. Sage, 1998
2. Goldstein H. **Multilevel Statistical Models**. 4th ed., Wiley, 2010
3. INVALSI **Le prove del Servizio Nazionale di Valutazione 2008-2009. Analisi tecnica**. 2009 [Available at: http://www.invalsi.it/download/2010/SNV_Prove_2008_2009.pdf; Retrieved: 11 December 2011]
4. Liu J., Dey D.K. **Skew random effects in multilevel binomial models: an alternative to nonparametric approach**. Statistical Modelling 8, 221-241, 2008
5. Petracco-Giudici M., Vidoni D., Rosati R. **Compositional effects in Italian primary schools: an exploratory analysis of INVALSI SNV data and suggestions for further research**. In: Organizational, Business, and Technological Aspects of the Knowledge Society – Communications in Computer and Information Science, Vol. 112, 460-470, Springer Berlin Heidelberg, 2010
6. Rabe-Hesketh S., Skrondal A. **Multilevel and Longitudinal Modelling Using Stata**. 2nd ed., Stata Press, 2008
7. Sani C. **Valutazione degli apprendimenti degli studenti della scuola primaria italiana: un'analisi multilivello**. Dissertation for degree in Statistical Sciences. University of Florence, 2011
8. Snijders T.A.B., Bosker R.J. **Multilevel analysis: an introduction to basic and advanced multilevel modelling**. 2nd ed., Sage, 2011
9. Stata Corp. **Stata 11 Reference Manual**, Stata Press, 2009

[1] Claudia Sani graduated Statistics in 2008. She got their MSc in Statistical Science in 2011 at the Department of Statistics, University of Florence. Her main areas of interest are Multilevel Analysis and Social and Medical Statistics. She is currently working for IMS Health Spa as Data Analyst.

[2] Leonardo Grilli is Associate Professor of Statistics at the University of Florence – Department of Statistics. His research concerns random effects models for multilevel analysis and methods of causal inference and effectiveness evaluation based on potential outcomes. The methodological work is accompanied by applications in socio-economic and other settings, particularly in effectiveness evaluation of educational systems.