The Annals of Applied Statistics
2016, Vol. 10, No. 4, 2405–2426
DOI: 10.1214/16-AOAS988

EXPLOITING TIMSS AND PIRLS COMBINED DATA: MULTIVARIATE MULTILEVEL MODELLING OF STUDENT ACHIEVEMENT[1]

BY LEONARDO GRILLI*, FULVIA PENNONI†, CARLA RAMPICHINI* AND ISABELLA ROMEO‡

University of Florence, University of Milano-Bicocca† and Mario Negri‡*

We illustrate how to perform a multivariate multilevel analysis in the complex setting of large-scale assessment surveys, dealing with plausible values and accounting for the survey design. In particular, we consider the Italian sample of the TIMSS&PIRLS 2011 Combined International Database on fourth grade students. The multivariate approach jointly considers educational achievement in Reading, Mathematics and Science, thus allowing us to test for differential associations of the covariates with the three outcomes, and to estimate the residual correlations among pairs of outcomes within and between classes. Multilevel modelling allows us to disentangle student and contextual factors affecting achievement. We also account for territorial differences in wealth by means of an index from an external data source. The model residuals point out classes with high or low performance. As educational achievement is measured by plausible values, the estimates are obtained through multiple imputation formulas.

1. Introduction. The role of large-scale assessment surveys in the public debate about education has dramatically grown since the mid-1980s. Despite the inevitable criticism, international achievement testing has the merit to display the great variability of the educational systems across the world and to shed light on the process underlying the growth of the human capital. As discussed in Rutkowski, von Davier and Rutkowski (2014), international achievement testing in education has many and ambitious purposes, including the assessment of policies and practices. Indeed, understanding the determinants of achievement in compulsory school is extremely important to design interventions at any level; see, among others, Reeve and Jang (2006) and the references therein.

In this paper we consider the large-scale assessment surveys TIMSS (Trends in International Mathematics and Science Study) and PIRLS (Progress in International Reading Literacy Study) by focusing on the Italian data. These surveys are generally carried out at different years; in 2011 for the first time the two cycles

Received March 2016; revised July 2016.

[1]Supported in part by the Grant "*Finite mixture and latent variable models for causal inference and analysis of socio-economic data*" (FIRB—Futuro in ricerca) funded by the Italian Government (RBFR12SHVV).

Key words and phrases. Hierarchical linear model, large-scale assessment data, multiple imputation, plausible values, school effectiveness, secondary data analysis.

coincided, thus providing a sample of students with a joint assessment in Reading, Math and Science. Italy represents an interesting case since it has a central educational system, whose egalitarian purposes are hampered by marked territorial differences in wealth.

In official reports, for any country the outcomes in Reading, Math and Science are analyzed separately by means of multilevel models [Foy and O'Dwyer (2013), Martin and Mullis (2013)]. We propose a multivariate multilevel approach, where the three scores are treated as a multivariate outcome measured for each student (level 1), where students are nested within classes (level 2). This approach allows us to gain further insights with respect to the univariate analysis, as we can estimate the residual correlations between pairs of outcomes at both hierarchical levels, which is important to make a comprehensive picture of student achievement and educational effectiveness. Moreover, a multivariate model enables to test whether the coefficient of an explanatory variable is identical across outcomes, for example, whether gender differences in achievement are the same for Reading and Math. From a methodological point of view, the multivariate multilevel model is a well established tool [Yang et al. (2002)], however, its potentialities have not yet been exploited in the framework of large-scale assessment data. In the following, we tackle some technical issues arising in this framework, such as the way of handling *plausible values* and sampling weights, and we discuss the model specification and the interpretation of the results.

The exploratory analysis on the Italian data shows that, for each outcome, the proportion of variability between classes is relevant, thus calling for an analysis of student characteristics (e.g., gender and family background), as well as contextual factors (e.g., school resources and wealth of the surrounding area). To this end, we exploit variables of the TIMSS&PIRLS combined dataset, with the addition of an external measure of wealth, namely the Gross Value-Added (GVA) at the province level. These variables allow us to adjust for prior differences among students and contexts. In the literature there is no general agreement on which contextual factors to adjust for: in principle, one should consider all the factors out of the control of the teachers or the school management, but the distinction is not always clear since a factor may be only partially out of control [Tekwe et al. (2004)]. In addition, the choice to adjust for a factor can be dictated by policy targets [Ladd and Walsh (2002)]. Given our interest in the methods, we do not tackle these issues, and we choose the covariates according to mainstream research and exploratory data analysis, as explained in Section 3. After adjusting for the selected covariates, the random effects collect unobserved contextual factors affecting student achievement, thus they could be interpreted in terms of comparative effectiveness, even if they do not have a causal meaning since students are not randomly allocated into classes. Keeping in mind the above limitations, we exploit predicted random effects to point out anomalous situations and territorial patterns. This analysis is relevant in an educational system which aims to be egalitarian, like the Italian one.

The paper is organized as follows. Section 2 describes the TIMSS&PIRLS 2011 survey, and then Section 3 focuses on the Italian sample, showing preliminary analyses and discussing the choice of covariates. Section 4 outlines the multivariate multilevel model, and then Section 5 illustrates the model selection process, reporting the main findings. Section 6 discusses alternative model specifications and deals with issues related to sampling weights. Finally, Section 7 gives final remarks and directions for future work.

2. The TIMSS&PIRLS 2011 survey. The large-scale assessment surveys TIMSS and PIRLS are organized by the International Association for the Evaluation of Educational Achievement (IEA). Specifically, TIMSS is an international assessment of mathematics and science achievements at fourth and eighth grades conducted every four years since 1995, whereas PIRLS provides information on trends in reading literacy achievement of fourth grade students every five years since 2001.

In TIMSS and PIRLS the students are selected by a complex multi-stage sampling design [Martin and Mullis (2012)]. The variables are obtained through questionnaires administered to students, their parents, their teachers and their school principals. The questionnaires of the two surveys are identical, except for subject-specific issues. For example, questions about teaching math are specific to the TIMSS teacher questionnaire. Parents completed the home questionnaire with questions about the child (e.g., literacy- and numeracy-centered activities at an early age), the family (e.g., home resources) and the parents themselves (e.g., level of education and employment status). The choice to collect the above information from the parents increases the quality of the derived variables as compared to surveys collecting the information from the student, where the responses may be seriously affected by the socio-economic status of the student [Jerrim and Micklewright (2014), Kreuter et al. (2010)].

In 2011 for the first time the TIMSS and PIRLS cycles coincided, enabling IEA to release the Combined TIMSS&PIRLS 2011 International Database including fourth grade students responding to both surveys. In the combined database the two surveys are perfectly comparable since they share the methodological framework and they are administered to the same sample of students. Indeed, IEA released the data as if they were collected by a single survey and created additional contextual scales by combining information from the two surveys. For example, the PIRLS scale "instruction affected by reading resource shortages" and the TIMSS scales "instruction affected by mathematics resource shortages" and "instruction affected by science resource shortages" were combined into a new contextual scale labelled "instruction affected by any resource shortages".

The TIMSS&PIRLS 2011 database provides achievement results scaled together in a multi-dimensional IRT model in order to preserve the correlation structure across the three achievement scales. Separate achievement scales are produced for Reading, Math and Science with an international mean of 500 points and a

standard deviation of 100 points, considering the 32 countries that administered the TIMSS and PIRLS 2011 assessments at the fourth grade [Foy (2013)]. Due to the joint scaling procedure, the scores in the combined dataset differ from those reported in the TIMSS 2011 and PIRLS 2011 reports.

For each achievement scale, five predictions of the student score, known as *plausible values* (PV), are provided. The variability among plausible values accounts for the uncertainty inherent in the scale estimation process [Martin and Mullis (2012), Mislevy (1991), Rutkowski et al. (2010), Wu (2005)]. Plausible values are not suitable as individual scores [von Davier, Gonzalez and Mislevy (2009)], but they are needed to account properly for the variability in the estimates for groups of students. In the following we exploit the five plausible values by using Multiple Imputation (MI) formulas [Rubin (2002), Schafer (2003)]. These formulas yield correct standard errors accounting for both the variability in imputing the scores and the variability in estimating the quantities of interest.

3. Preliminary analysis and choice of the variables. Our analysis concerns Italy, where 4200 students participated in TIMSS and 4189 participated in PIRLS. The corresponding TIMSS&PIRLS 2011 combined dataset includes 4125 students who responded to both surveys, thus having plausible values for all the outcomes (Reading, Math and Science). Note that combining the surveys caused a negligible reduction of the sample size. The students are nested in 239 classes, which are nested in 202 schools. Italian primary schools belong to a public system: the majority of schools are operated by the state and the other ones must still adhere to common guidelines about study programs. Generally, pupils attend the nearest school from home. Even if the parents can choose a different school according to the availability of places, this is uncommon except in large cities.

First of all we perform a descriptive analysis of the sample without using sampling or adjusting weights. Exploiting the five plausible values, the average scores for Italy, alongside with their within-PV standard deviations (in parentheses), are as follows: Reading 525 (75.5), Math 502 (76.4) and Science 518 (77.6). The above means for the outcomes differ from the means reported in the stand-alone TIMSS 2011 and PIRLS 2011 reports, especially for Reading. The main reason of the discrepancies is that the scores in the combined dataset are computed using a joint scaling procedure [Foy (2013)].

Table 1 reports sample sizes and summary statistics on achievement for Italy by geographical area, showing a decrease in average scores moving from North to South, whereas the standard deviations among classes have an opposite tendency. Despite the public system, student achievement is markedly different across geographical areas; indeed, the range of the average score is close to the standard deviation among classes. The geographical pattern of achievement reflects some well-known differences in the economic well-being of the country. Student achievement is associated with wealth mainly because of the relationship with the socio-economic condition of the area, which is a key factor in the literature on

TABLE 1

TIMSS&PIRLS 2011 *sample sizes and average Gross Value Added* 2010 *for Italy by geographical area, alongside with MI combined average scores and their MI combined between-class standard deviations* (*in parenthesis*)

	Sample sizes				Average score (between-class SD)		
Area	**Classes**	**Teachers**	**Students**	**GVA**	**Reading**	**Math**	**Science**
North-West	48	103	849	122	540 (23)	516 (28)	538 (28)
North-East	49	103	920	120	534 (38)	508 (42)	529 (43)
Centre	48	97	852	113	530 (31)	506 (35)	524 (36)
South	49	97	832	66	517 (40)	499 (54)	508 (53)
South–Islands	45	83	672	69	502 (48)	475 (57)	489 (58)
Italy	239	483	4125	100	525 (39)	502 (46)	518 (47)

school effectiveness [Hanushek and Woessmann (2011), Stonge, Ward and Grant (2011)]. In particular, some auxiliary services are provided by the municipalities, thus schools in wealthier areas may benefit from richer extracurricular activities and better services, such as canteen and bus services.

In TIMSS and PIRLS, the wealth of the surrounding area of the school is measured through some questions posed to the school principal, but they are of little value since the extent of the area to be considered is undefined, and the judgement has a subjective nature. We thus prefer to rely on an external measure of wealth, namely, the per-capita Gross Value Added (GVA) at market prices in 2010 [Istituto Tagliacarne (2011)]. This index is defined at the province level, which is the finest geographical level available for a nationwide wealth index. We recognize that GVA neglects the within-province heterogeneity, which can be substantial. Notwithstanding, we argue that GVA is the best available measure for the purpose of adjusting school effectiveness for wealth.

The GVA is measured for each of the 110 Italian provinces, ranging from 55 to 142, with 100 representing the national average. Figure 1 shows the patterns at the province level of GVA (left panel) and the average score on Math (right panel, where white areas correspond to provinces without sampled schools). Both quantities tend to decrease from North to South, even if the average score on Math is more irregular. The relationship between Math score and GVA is also represented in Figure 2 through a local polynomial smoothing: achievement is positively related to wealth, even if the relationship is weak and it holds only for provinces with a GVA below the Italian average of 100. Reading and Science have similar relationships.

In order to adjust the achievement scores for student and contextual factors, we consider a subset of the variables available in the combined dataset. The choice of this subset is driven by theoretical arguments [Hanushek and Woessmann (2011)] and previous studies [Chiu and Xihua (2008), Hammouri (2004), Wang, Osterlind and Bergin (2012)]; in particular, we rely on the technical Appendix B of the

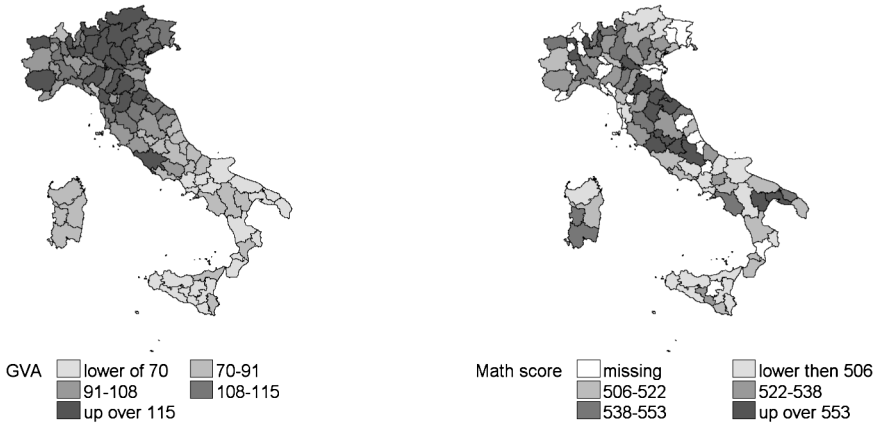

FIG. 1. *Cartograms by province of the Gross Value Added* 2010 (*left panel*) *and the Math average score from TIMSS&PIRLS* 2011 (*right panel*).

TIMSS&PIRLS 2011 Report [Foy and O'Dwyer (2013)]. In addition, for the suggested covariates we perform an exploratory analysis to assess their role in our case study.

Descriptive statistics of the selected covariates are shown in Table 2. The first column reports the sample sizes at the relevant levels (student, teacher, class, school), whereas the last column reports the number of observations for each variable: in most cases there are missing values, though their percentage is small (at most 8.6%).

At the student level, we include dummy variables for gender (1 if female), preschool (1 if the student attended at least 3 years) and language spoken at home (1 if not Italian). Furthermore, two home background scales from TIMSS&PIRLS

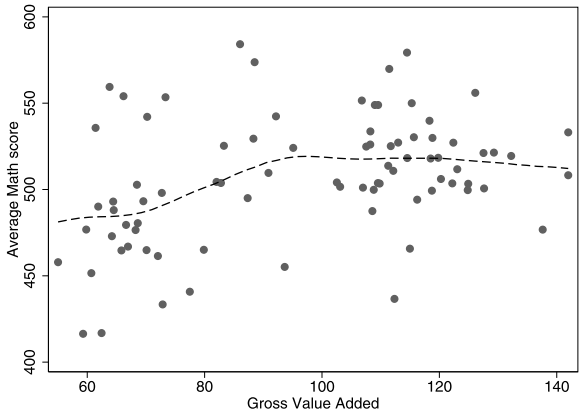

FIG. 2. *Local polynomial smoothing of the province average of Math score* (*TIMSS&PIRLS* 2011) *as a function of province Gross Value Added* (2010).

TABLE 2
Summary statistics of selected variables by hierarchical level (*sample sizes in parenthesis*, *TIMSS&PIRLS* 2011, *Italy*)

Variable	**Mean**	**SD**	**Min.**	**Max.**	***N***
Student level (4125)					
Female	0.51	–	0	1	4125
Preschool	0.75	–	0	1	3826
Language at home not Italian	0.21	–	0	1	4083
Home resources for learning	9.72	1.55	3.41	15.29	3770
Early literacy/numeracy tasks	9.24	1.60	3.65	12.90	3846
Teacher level (483)					
Woman	0.96	–	0	1	458
Age					458
<40	0.15				
40–49	0.38				
≥50	0.47				
Degree	0.35	–	0	1	455
Years of teaching	23.34	10.29	1	42	446
Subject(s) taught					483
Reading only	0.35				
Mathematics and Science	0.28				
Mathematics only	0.12				
Science only	0.10				
Mathematics, Science and Reading	0.06				
Science and Reading	0.05				
Mathematics and Reading	0.03				
Class level (239)					
Class Mean (CM) of "Female"	0.51	0.13	0	1	239
CM of "Preschool"	0.69	0.17	0	1	239
CM of "Language at home not Italian"	0.23	0.17	0	1	239
CM of "Home resources for learning"	9.64	0.83	7.17	11.90	238
CM of "Early literacy/numeracy tasks"	9.24	0.51	7.58	10.89	238
School level (202)					
Adequate environment and resources	9.62	1.07	7.13	13.62	201
School is safe and orderly	9.41	0.88	7.36	12.37	202
School with Italian students >90%	0.65	0.48	0	1	196
Less than 10% students with low SES	0.39	0.49	0	1	191
School is located in a big area	0.34	0.48	0	1	198
In the area live more then 50,000 people	0.28	0.45	0	1	196
Six days of school per week	0.47	0.50	0	1	198

2011 are used to describe the student home environment: *Home Resources for Learning* and *Early Literacy/Numeracy Tasks*, described in detail by Martin and Mullis (2013). In summary, *Home Resources for Learning* is derived from items on the number of books and study supports available at home and parents' levels

of education and occupation; on the other hand, *Early Literacy/Numeracy Tasks* is the student average score on two scales derived from the parents' responses on how well their child could do some early literacy and numeracy activities when beginning primary school.

In general, the assessment of school effectiveness should adjust for prior achievement, for example, the level of achievement at the beginning of the school cycle. TIMSS&PIRLS data do not contain any direct measure of prior achievement, but this is a minor limitation since (i) the primary school is the first compulsory cycle, thus the programs start with the basics of each discipline, and (ii) we include in the model some variables related to the skills of the pupil acquired before primary school, namely, the scale *Early Literacy/Numeracy Tasks* and *Preschool*, an indicator for having attended at least three years of preschool. Information about attending preschool, defined as ISCED level 0 (from 0 to 5 years), is provided by the parents. We consider the indicator of attending at least three years of preschool (75% in the sample), since this category includes pupils who completed the standard cycle of pre-primary education.

At teacher level we consider gender (1 if woman), age group, education (1 if the teacher has a degree) and years of teaching. Note that the average number of years of teaching is rather high (23.34) and the percentage of teachers with a university degree is only 35%: indeed, until 2002 a special purpose high school diploma was enough to become a teacher in primary school. In any class, every subject is taught by a single teacher, but a teacher may be in charge of one or more subjects. Table 2 shows that Reading usually has a specific teacher, whereas Math and Science are often taught by the same teacher, even if other combinations of teachers and subjects are also possible. The class-level variables in Table 2, defined as averages of the corresponding student-level covariates, are intended to capture contextual effects, including peer effects [Ammermueller and Pischke (2009)]. At school level, *Adequate environment and resources* and *School is safe and orderly* are contextual scales [Martin and Mullis (2013)], while the other school variables in Table 2 are directly based on the answers of school principals.

4. Model specification. TIMSS&PIRLS data have a hierarchical structure, with students nested in classes, which are nested in schools. Moreover, given our interest in territorial patterns, provinces can be considered as a fourth level. However, as illustrated in Section 6, the data under consideration do not provide enough information to fit a four-level model. Therefore, we fit a two-level model with classes at level 2. In this way, the random effects collect all the unobserved contextual factors at class and higher hierarchical levels. In general, ignoring top hierarchical levels inflates the variance component at the highest level of the specified model. Indeed, this variance component collects all the sources of variance at higher levels [Tranmer and Steel (2001)]. We adjust for the correlation between classes of the same school by using robust standard errors for clustered observations [Rabe-Hesketh and Skrondal (2006)]. Note that our approach is different

from the one adopted in official reports [Foy and O'Dwyer (2013)], where multilevel models are specified with schools as level 2 units.

In the following we outline the notation for the two-level model used in our analysis, which is a multivariate multilevel model [Goldstein (2011), Snijders and Bosker (2012), Yang et al. (2002)]. Alternative specifications about the distributions of the errors are discussed in Section 6. Let Y_{mij} be the score on the mth outcome for the ith student of the jth class, with $m = 1, 2, 3$ (1: Reading, 2: Math, 3: Science), $i = 1, \ldots, n_j$, $j = 1, \ldots, J$. The number of students of the jth class is denoted with n_j, whereas the total number of students is denoted with $N = \sum_{j=1}^{J} n_j$. The Italian sample of the TIMSS&PIRLS 2011 Combined Dataset includes $N = 4125$ students nested into $J = 239$ classes. We specify the following multivariate two-level model for outcome m of student i in class j:

$$Y_{mij} = \alpha_m + \boldsymbol{\beta}_m' \mathbf{x}_{mij} + \boldsymbol{\gamma}_m' \mathbf{w}_{mj} + u_{mj} + e_{mij}, \tag{1}$$

where $\mathbf{x}_{mij}$ is the vector of student covariates (level 1), and $\mathbf{w}_{mj}$ is the vector of contextual covariates (level 2), including variables measured at the level of class, school or province. All the vectors have the outcome index m since they can include outcome-specific covariates, such as the characteristics of the teacher. Level 1 errors e_{mij} are assumed independent across students. Level 2 errors (random effects) u_{mj}, which collect all the unobserved contextual factors for outcome m, are assumed independent across classes and independent from level 1 errors.

We make standard assumptions for the distributions of the model errors, including homoscedasticity (within each outcome) and normality. Specifically, level 1 errors $\mathbf{e}_{ij}' = (e_{1ij}, e_{2ij}, e_{3ij})$ are assumed to be multivariate normal with zero means and variance–covariance matrix

$$\operatorname{Var}(\mathbf{e}_{ij}) = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ & \sigma_2^2 & \sigma_{23} \\ & & \sigma_3^2 \end{pmatrix}, \tag{2}$$

whereas level 2 errors $\mathbf{u}_j' = (u_{1j}, u_{2j}, u_{3j})$ are assumed to be multivariate normal with zero means and variance–covariance matrix

$$\operatorname{Var}(\mathbf{u}_j) = \mathbf{T} = \begin{pmatrix} \tau_1^2 & \tau_{12} & \tau_{13} \\ & \tau_2^2 & \tau_{23} \\ & & \tau_3^2 \end{pmatrix}. \tag{3}$$

Therefore, the response vector $\mathbf{Y}_{ij} = (Y_{1ij}, Y_{2ij}, Y_{3ij})'$ has a residual variance–covariance matrix $\operatorname{Var}(\mathbf{Y}_{ij}) = \boldsymbol{\Sigma} + \mathbf{T}$.

5. Model selection and results. The analysis is based on the multivariate two-level model of equations (1)–(3) fitted by maximum likelihood. In order to account for the variability induced by plausible values, estimation is performed separately for each of the five plausible values, and then the results are combined

by Multiple Imputation (MI) formulas [Rubin (2002), Schafer (2003)]. These formulas yield correct standard errors accounting for both the variability in imputing the scores and the variability in estimating the model parameters. The analysis is carried out by using the `mixed` and `mi` commands of Stata [StataCorp (2013)].

The estimation sample consists of 3741 students and 237 classes. Indeed, 384 students (9.3%) and 2 classes (0.8%) have been excluded due to missing values in the covariates (see Table 2). Among the classes in the estimation sample, only 8 (3.3%) lose more than one-third of their pupils due to missing values in the covariates. In our application, missing values are rare at class and school levels. At the student level, the covariates with the highest percentage of missing values are those derived from the parents questionnaire, namely, *Preschool* (7.2%), *Home resources for learning* (8.6%) and *Early literacy/numeracy tasks* (6.8%).

An alternative to deleting records with missing values is represented by multiple imputation methods, which have recently been extended to complex multilevel settings [Goldstein, Carpenter and Browne (2014)]. In the framework of large-scale assessment data, imputation methods have been considered by Bouhlila and Sellaouti (2013), Foy and O'Dwyer (2013), Weirich et al. (2014). In our data, missing values mostly concern the covariates derived from the parents questionnaire. Unfortunately, those covariates tend to be missing altogether, thus it is difficult to specify an effective imputation model. The missing mechanism seems to be related to the family background, and so deleting students with missing values may lead to some bias in the estimates. This kind of bias is likely to remain still after imputing missing values [Rubin (2002)]. Weighting the costs and benefits, we decide to not impute the missing values.

5.1. *Model selection*. The model selection procedure in principle involves fitting the multivariate multilevel model repeatedly, each time combining the estimates with MI formulas. In order to speed up the selection process, we adopt two simplifications: (i) the outcomes are analyzed separately with univariate multilevel models, retaining covariates being significant in at least one of the univariate models, and (ii) estimation is carried out using only the first plausible value. Using a single plausible value gives underestimated standard errors, implying a conservative selection of the covariates.

In order to test for contextual effects of level 1 covariates, we include in the model the corresponding level 2 means, namely, the class-level averages reported in Table 2. Contrary to official reports [Foy and O'Dwyer (2013)], we do not center level 1 covariates at their level 2 means so that the coefficient of a level 2 mean is interpreted as a contextual effect, namely, the difference among the between effect and the within effect [Raudenbush and Willms (1995)]. In order to enhance the interpretability of the intercept, we center continuous covariates at their sample grand means, except for GVA which is centered at 100 (Italian average).

According to the usual model selection strategy [Snijders and Bosker (2012)], we add covariates following the data hierarchical structure, namely, we first specify the level 1 model introducing student covariates, and then we add higher level

TABLE 3
Multivariate multilevel model: *main steps of the model selection process*, *first plausible value* (*TIMSS&PIRLS* 2011, *Italy*)

Model	n. par.	log L	Significant covariates (on at least one subject)
M0: null	15	−59,625.44	–
M1: student covariates	30	−59,119.09	Female, Preschool, Language spoken at home is not Italian, Home resources for learning, Early literacy/numeracy tasks
M2: student and class/school covariates	36	−59,109.75	M1 covariates + Class average Early literacy/numeracy tasks, School adequate environment and resources
M3: student, class/school and province covariates	36	−59,106.25	M1 covariates + School adequate environment and resources, Gross Value Added by province

covariates (class, school, province). We also consider teacher characteristics, however, none of them are significant. Table 3 reports the models selected at the end of each step. All the considered models have 12 variance–covariance parameters: 6 for the within-class covariance matrix in equation (2) and 6 for the between-class covariance matrix in equation (3). Note that models *M2* and *M3* have the same number of parameters since, after the inclusion of GVA, the class mean of *Early literacy/numeracy tasks* is no more significant.

5.2. *Results from the null model*. The model without covariates allows us to explore the correlation structure of the three outcomes. The upper part of Table 4 summarizes the results of the null model in terms of correlation matrices and between-class proportions of variances and covariances after the application of MI formulas. The *within-class* and *between-class* correlation matrices are derived from the corresponding covariance matrices $\mathbf{\Sigma}$ and $\mathbf{T}$ of equations (2) and (3), whereas the total correlation matrix is derived from the total covariance matrix $(\mathbf{\Sigma} + \mathbf{T})$. Table 4 shows that the three scores are highly correlated, in particular at level 2.

The rightmost matrix in Table 4 reports the percentage of variances and covariances at level 2, namely, each element of $\mathbf{T}$ is divided by the corresponding element of $(\mathbf{\Sigma} + \mathbf{T})$. For example, the percentage of level 2 variance for Reading is $100 \times \hat{\tau}_1^2/(\hat{\sigma}_1^2 + \hat{\tau}_1^2) = 19.8$, which is also known as the ICC (Intraclass Correlation Coefficient). Note that Reading is the subject with the lowest ICC, maybe because it is most influenced by student background characteristics.

The estimated ICCs show that contextual factors explain a relevant portion of the variability in achievement. These values are in line with the school-level ICCs in TIMSS&PIRLS reports [Martin and Mullis (2013)]; with respect to the countries participating in TIMSS&PIRLS, the ICCs for Italy are intermediate.

TABLE 4
Correlation matrix decomposition and between-class percentage of (co)variances. Estimates from null model (M0) and final model (M3), MI combined estimates (TIMSS&PIRLS 2011, Italy)

	Correlations									% Between-class (co)variances		
	Within-class			Between-class			Total					
	Read	**Math**	**Scie**	**Read**	**Math**	**Scie**	**Read**	**Math**	**Scie**	**Read**	**Math**	**Scie**
M0 null												
Reading	1.00			1.00			1.00			19.8		
Math	0.71	1.00		0.93	1.00		0.76	1.00		29.5	28.8	
Science	0.81	0.74	1.00	0.97	0.98	1.00	0.85	0.81	1.00	28.2	35.0	29.4
M3 final												
Reading	1.00			1.00			1.00			16.3		
Math	0.67	1.00		0.93	1.00		0.72	1.00		27.6	27.6	
Science	0.77	0.70	1.00	0.97	0.97	1.00	0.80	0.78	1.00	25.3	34.1	26.9

5.3. *Results from the final model.* The results from the final model are obtained by fitting model *M3* of Table 3 separately for each plausible value and then combining the estimates through MI formulas. The lower part of Table 4 summarizes the results in terms of residual correlation matrices and residual between-class percentage of variances and covariances. These matrices are similar to those of the null model (upper part of the table), with slight reductions of within-class correlations and between-class percentages.

Table 5 reports the estimates and the robust standard errors for regression coefficients and variance–covariance parameters. With reference to the considered covariates summarized in Table 2, all the student-level covariates are significant, but the corresponding class means are not. None of the teacher covariates are significant. At the school level, the only significant variable is *Adequate environment and resources*. At the province level, GVA is significant.

The last column of Table 5 reports the p-value of the F test for the equality of the regression coefficients across the three outcomes. For example, for the sth student-level covariate, the null hypothesis is $H_0 : \beta_{1s} = \beta_{2s} = \beta_{3s}$. The test, which is feasible only for a multivariate model, is performed by using the procedure `mi testtr` of Stata, implementing formula (1.17) of Li et al. (1991). Interestingly, for all the contextual covariates the magnitude of the association with the three outcomes is similar, while the student-level covariates show different relationships with the outcomes, except for preschool. Also, note that family background covariates have a similar association with Reading and Science, as opposed to Math, and therefore the abilities required for Science seem to be closer to those for Reading than to those for Math. Likely, this is a consequence of the fact that, in the Italian primary schools, the approach to Science is mainly qualitative, thus reading ability is more important than math ability.

TABLE 5

Multivariate multilevel model: *parameter estimates and robust standard errors from the final model, MI combined results* (*TIMSS&PIRLS* 2011, *Italy*)

	Read		**Math**		**Science**		***F* test**[†]
	Coef.	**s.e.**	**Coef.**	**s.e.**	**Coef.**	**s.e.**	***p*-value**
Intercept	531.73	3.57	514.99	4.25	531.47	3.92	0.0006
Student covariates							
Female	2.92	2.41	−11.96	3.05	−10.64	2.28	0.0000
Lang. home not Italian	−22.57	3.12	−14.94	3.27	−23.74	3.53	0.0161
Preschool	8.85	3.01	8.46	2.51	10.91	3.15	0.6386
Home res. for learning	14.04	0.84	10.64	0.84	13.23	0.93	0.0009
Early lit./num. tasks	7.24	0.77	10.07	0.76	6.53	0.83	0.0051
School covariates							
Adequate envir. & res.	5.28	1.92	8.61	3.19	7.00	2.96	0.1950
Province covariates							
GVA (below 100)	0.45	0.15	0.48	0.21	0.55	0.20	0.3983
Between-class cov. matrix							
Variances: τ_m^2	725.7	192.0	1332.3	225.1	1274.1	262.1	
Cov (Read, Math): τ_{12}	915.2	195.9					
Cov (Math, Scie): τ_{23}			1266.1	234.7			
Cov (Read, Scie): τ_{13}					931.6	221.1	
Within-class cov. matrix							
Variances: σ_m^2	3716.1	101.9	3500.1	120.8	3471.7	132.6	
Cov (Read, Math): σ_{12}	2400.9	86.0					
Cov (Math, Scie): σ_{23}			2452.3	91.6			
Cov (Read, Scie): σ_{13}					2757.4	105.9	

[†] F test for the equality of regression coefficients among the three outcomes.

The intercepts in Table 5 represent the average scores for the baseline student, which is a male, whose language spoken at home is Italian, who did not attend at least three years of preschool, and with all the other covariates set at their mean. The performance of the baseline student is beyond the international mean of 500 in all the considered outcomes, though the average score in Math is substantially lower than the average scores in Reading and Science. According to the F test, this difference is significant.

The regression coefficients are significant for all the considered outcomes, except for being female, which is not significantly associated with Reading.

In general, the coefficients have the expected signs. Females have a lower performance in Math and Science, but not in Reading. Students from families not speaking Italian at home have a lower performance in all the subjects, especially in Reading and Science. Students who attended preschool for at least three years have a better performance, with no significant difference among the three outcomes. The two home background questionnaire scales are positively related with student

achievement. However, *Home Resources for Learning* (including number of books at home and education level and employment status of parents) has a greater association with Reading and Science, while *Early Literacy/Numeracy Tasks* (measuring how well the child could do several early literacy and numeracy activities when beginning primary school) has a stronger association with Math. Thus, achievements in Reading and Science are more related to cultural and socio-economic factors of the family, while achievement in Math is more related to specific activities in early childhood.

At the school level, disposing of *Adequate Environment and Resources* is associated to a higher score, with no significant difference across outcomes. The socio-economic context of the province where the school is located is measured by the GVA index. On the basis of the local polynomial smooth of Figure 2, the relationship of achievement with GVA is modelled by a linear spline with a single knot in 100 (the national average). Consistently with the relationship highlighted in Figure 2, the line for GVA < 100 has a significant positive slope. On the contrary, the line for GVA > 100 is nearly flat and the slope is not significantly different from zero, and thus we constrain such slope to zero. In this way, wealth is associated with the student achievement only for provinces with GVA below the national average. The regression coefficient of GVA is similar across outcomes, and it amounts to about a half point in the score for each point in the index. The province with the lowest GVA is 45 points below the national average so that the maximum reduction in achievement associated to GVA is about -22.5 points.

The proportions of variance explained by the final model with respect to the null model are higher at level 2. Indeed, the within-class variances reduce by about 15% for the three outcomes, whereas the between-class variances reduce by 33% for Reading, 20% for Math and 26% for Science. The reduction of the between-class variances is due to the contextual variables at school and province levels and to the compositional effects of student background covariates. Such compositional effects capture cultural and socio-economic factors that are more related to the achievement in Reading, whose level 2 residual variance shows the greatest reduction. Even if the reduction of variances is stronger at level 2, the residual ICCs derived from the estimated variances of Table 5 are quite high, specifically 16.3% for Reading, 27.6% for Math and 26.9% for Science. These values point out the existence of relevant unobserved contextual factors. The correlations among outcomes are similar to those observed in the null model, reported in Table 4. In particular, the estimated correlations among the level 2 errors of the three outcomes are very high, reaching at least 0.93.

5.4. *Analysis of predicted random effects*. The main aim of TIMSS and PIRLS surveys is to perform international comparisons in terms of student achievement, and thus official reports publish country unadjusted averages and exploit multilevel models to understand individual and contextual determinants of achievement

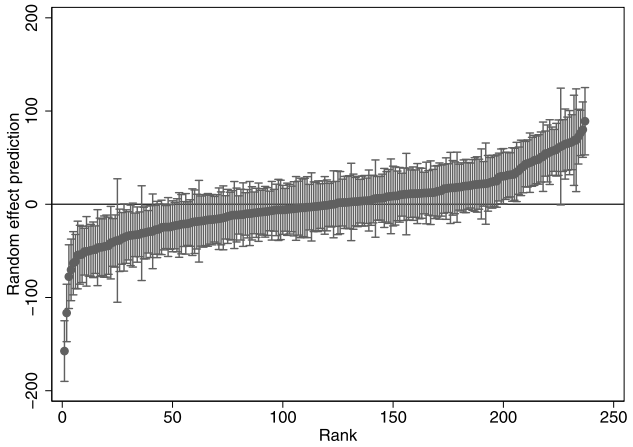

FIG. 3. *Empirical Bayes predictions of the random effects for Math with* 95% *confidence intervals, MI combined results* (*TIMSS&PIRLS* 2011, *Italy*).

[Foy and O'Dwyer (2013)]. In this section, we use the multilevel model for a different purpose, namely, monitoring the educational system of a single country in terms of comparative effectiveness of the classes. This analysis also allows us to detect institutions or areas with especially good or poor performances. To this end, we exploit predictions of the random effects (level 2 errors) of the multivariate multilevel model (1), which represent the contribution of unobserved contextual factors to student achievement, adjusted for differences in student characteristics and observed contextual factors. These predictions can be interpreted as measures of comparative effectiveness, as discussed in Section 1.

The analysis of predicted random effects is carried out by combining the results derived from the models fitted on the five plausible values. For each plausible value, once the model has been fitted via maximum likelihood, the level 2 error u_{mj} of outcome m for class j is predicted by the Empirical Bayes (EB) method, yielding a prediction and its standard error [Goldstein (2011), Snijders and Bosker (2012)]. These predictions are combined by standard MI formulas.

As the residual correlations among the three outcomes are very high (above 0.93), in the following we illustrate the analysis of the predicted random effects with reference to a single outcome, namely Math. Figure 3 shows a caterpillar plot where EB predictions are reported in increasing order and endowed with their 95% confidence intervals. This plot facilitates the comparison of classes in terms of effectiveness as defined in Section 1. We stress that classes at the extremes of the ranking for Math stay in a similar position for Reading and Science. For example, the two classes at the bottom of Figure 3 perform badly in all the considered disciplines.

In the caterpillar plot, a class whose confidence interval does not intersect zero has a degree of effectiveness significantly different from the population mean.

TABLE 6
Proportions of good *and* poor *classes based on EB residuals from the final model M*3 (*TIMSS&PIRLS* 2011, *Italy*)

Area	Classes	Proportion of classes	
		good	***poor***
North-West	48	0.167	0.083
North-East	49	0.102	0.082
Centre	47	0.043	0.213
South	49	0.286	0.224
South and Islands	44	0.227	0.250
Italy	237	0.165	0.169

Specifically, a class with an interval above zero has a *good* effectiveness, since its student average achievement is significantly higher than the level expected on the basis of the covariates. On the contrary, a class with an interval below zero has a statistically significant *poor* effectiveness. As for the 237 classes depicted in Figure 3, it turns out that 39 classes show a *good* level and 40 show a *poor* level of effectiveness.

The fitted model accounts for differences in wealth across provinces by means of the GVA index, and thus the residuals may reveal additional territorial differences not captured by GVA. Table 6 reports the proportions of *good* and *poor* classes by geographical area based on level 2 residuals: in North-West *good* classes prevail on *poor* classes, while in the Centre the pattern is reversed. This points out a residual territorial influence on mean achievement beyond GVA. However, if we account for this influence by adding geographical dummies in the fixed part of the model, then their coefficients turn out to be not significant. In the two Southern areas, the proportions of *good* and *poor* classes are higher than in the rest of Italy. This result confirms that schools in Southern regions have a higher variability in effectiveness, as found by Sani and Grilli (2011) using national standardized test data collected by the Italian Institute for the Evaluation of the Educational System (INVALSI). As mentioned in Section 6, such differential variability could be modelled through heteroscedastic random effects, but in the present application there is no significant improvement in the model log-likelihood.

6. Complementary issues.

6.1. *Alternative model specifications.* The data have a four-level hierarchical structure with pupils nested in classes, which are nested in schools, which are nested in provinces. In principle, the multivariate model (1) should have random effects for classes, schools and provinces. However, fitting the four-level version of the multivariate model yields computational problems as the Hessian matrix is

TABLE 7
Univariate four-level models without covariates: *variance decomposition* (*TIMSS&PIRLS* 2011, *Italy*)

Level	**Reading**		**Math**		**Science**	
Province	135.1	(2.5%)	182.5	(3.1%)	315.8	(5.3%)
School	640.2	(11.8%)	1004.2	(17.2%)	1127.5	(18.9%)
Class	238.9	(4.4%)	455.9	(7.8%)	308.7	(5.2%)
Pupil	4391.0	(81.2%)	4180.3	(71.8%)	4214.9	(70.6%)
Total	5405.2	(100.0%)	5823.0	(100.0%)	5966.9	(100.0%)

not positive definite. These problems are due to data limitations. Indeed, in the estimation sample there are 237 classes nested into 200 schools; thus most schools only have one class and there is little information on the variability of classes within schools.

Notwithstanding, the variance decomposition at the four levels can be estimated by fitting three univariate four-level models separately for each subject. The results are reported in Table 7. As expected, the largest part of variability is at the pupil level. The contextual factors are more relevant for Math and Science as compared to Reading, the largest contribution being at the school level. The province-level factors give a minor contribution to the variability (from 2.5% to 5.3%), which is nearly halved in the final model controlling for wealth through the GVA index. It is worth noting that in the two-level model used in the analysis discussed in Section 5, the level 2 variances and covariances collect the contribution from the unobserved contextual factors at class, school and province levels.

Other extensions of the model concern relaxing the standard assumptions on the errors [Grilli and Rampichini (2015)] outlined in Section 4. In particular, it is worth considering a specification with heteroscedastic random effects where the covariance matrix $\mathbf{T}$ of equation (3) varies according to geographical areas to account for differential variability [Sani and Grilli (2011)]. However, the heteroscedastic specification does not significantly improve the model fit, which is not surprising in light of the complexity of the considered model, where the random effects at level 2 are characterized by a 3×3 covariance matrix.

Another heteroscedastic specification could be devised to account for a possible differential variability related to the didactic organization of the class: indeed, 11% of the classes have a single teacher for the three subjects, 75% of the classes have two teachers, while the remaining classes have three teachers. The sample correlations among the outcomes are slightly higher in classes with a single teacher. This pattern can be accommodated by a model with two covariance matrices, one for classes with a single teacher and the other for classes with multiple teachers. However, this complex specification does not yield a significant improvement in the model fit.

6.2. *Weighted estimation*. The sampling scheme adopted by TIMSS and PIRLS is a stratified two-stage cluster sample design, with schools as primary units and classes as secondary units; all the students of the selected classes enter the sample [Martin and Mullis (2012)]. The released dataset contains weights separately for each hierarchical level see also Joncas and Foy (2013). At any level, the weight is defined as the product of the *sampling weight* (i.e., the reciprocal of the conditional sampling probability) and the *adjustment weight*, which accounts for nonparticipation of sampled units. The overall student weight is obtained by multiplying the weights across the three hierarchical levels (student i, class c, school s), namely, $w_{ics} = w_{i|cs} w_{c|s} w_s$.

In a regression model, weights are needed to obtain unbiased estimates when the sampling is informative, namely, the inclusion probabilities are related with the model errors, which is an assumption not directly verifiable. Unfortunately, sample weights inflate the standard errors of the estimators, and thus the trade-off between bias and variance should be evaluated case by case. The informativeness of the sampling design in multilevel modelling must be evaluated separately for each hierarchical level: given that our multilevel model (1) has students nested into classes, we consider the conditional student weights $w_{i|cs}$ and the unconditional class weights $w_j = w_{cs} = w_{c|s} w_s$. It is worth noting that the conditional student weights $w_{i|cs}$ are constant within the classes by construction, and thus they are not informative and they can be ignored in a model-based approach. In order to evaluate the informativeness of the unconditional class weights w_j, we apply some of the methods proposed in the literature [Rabe-Hesketh and Skrondal (2006), Rutkowski, von Davier and Rutkowski (2014), Snijders and Bosker (2012)]. First of all, we assess the variability of the weights by computing the *design effect* (*deff*): in the estimation sample, the number of classes is 237, while the effective sample size $(\sum w_j)^2 / \sum w_j^2$ is 198.8, and thus the design effect at level 2 is 198.8/237 = 0.84. This value is not much lower than one, and thus the potential bias due to ignoring the weights is limited. We further investigate this issue by comparing the estimates of the regression coefficients of the multilevel multivariate model (1) obtained with and without incorporating the level two weights w_j into the estimation algorithm (using the first plausible value for simplicity). As expected, weighting inflates most standard errors. The impact of weighting on the point estimates can be summarized, for instance, by the index $I_2 = |\hat{\beta}^w - \hat{\beta}^u| / \text{s.e.}(\hat{\beta}^u)$ suggested by Asparouhov (2006), where $\hat{\beta}^w$ and $\hat{\beta}^u$ are the weighted and unweighted estimates, respectively. In our case I_2 ranges from 0.07 to 0.58, pointing out a limited impact of weighting on the estimates [see also Snijders and Bosker (2012), page 240]. Therefore, considering the loss of efficiency, we carry out the analysis without weights.

7. Final remarks. We carried out a secondary data analysis of the Italian sample of the TIMSS&PIRLS 2011 Combined International Database. This database

provides an opportunity to perform, for the first time with TIMSS and PIRLS surveys, a joint analysis of achievement in Reading, Math and Science for fourth grade students. The analysis relies on a multivariate multilevel model, thus accounting for both the multivariate nature of the outcome and the hierarchical structure of the data.

The additional findings allowed by the multivariate approach are twofold. First, the multivariate model enables us to test for differences in the regression coefficients, thus pointing out differential effects of the covariates on the three outcomes. Notably, females have a lower performance in Math and Science, but not in Reading, whereas student background characteristics have a similar relationship with Reading and Science, as opposed to Math. On the other hand, contextual factors are associated in the same way with the three outcomes. Second, the multivariate model enables us to study the correlations among the outcomes, discovering that they are high at both student and class levels, even after adjusting for individual and contextual factors. In particular, the residual class-level correlations are so high that the three outcomes are indistinguishable in terms of effectiveness.

A peculiarity of our analysis lies in the use of the per capita Gross Value Added at the province level (GVA) as an indicator of territorial differences in wealth. The relationship between student achievement and GVA is well represented through a spline: the student achievement is positively related to wealth for provinces below the national average, with no significant relationship for provinces above the national average. Alternatively, differences among geographical areas could be accounted for by means of dummy variables, however, the GVA index has the merit of yielding a parsimonious model with an economic interpretation. The analysis of predicted random effects allows us to identify few classes with extremely high or low effectiveness and to investigate further patterns not described by the model, such as the higher variability in effectiveness among classes in the Southern regions.

The scaling methodology of TIMSS&PIRLS is similar to that of PISA, which has been criticized by several scholars [Goldstein (2004), Kirsch et al. (2002), Kreiner and Christensen (2014)]. However, we argue that scaling issues have a limited impact on our results since we consider a given grade at a fixed time in a single country. In particular, we do not make comparisons across countries. Notwithstanding, some concerns remain on the adequacy of the model used to generate the plausible values, namely the imputation model. In fact, the TIMSS&PIRLS documentation explains that the model is conditioned on a subset of the principal components of the background variables, in addition to a few "primary" variables such as student gender. The imputation model is not a multilevel one, and thus the standard MI formulas do not necessarily produce unbiased estimates of standard errors. The details on model specification and fit measures are not provided, and thus it is not possible to judge the adequacy of the imputation model. In principle, problems related to the imputation model can be avoided by specifying a multilevel IRT model based on item responses [Fox and Glas (2001), Johnson and

Jenkins (2005)]. However, due to the high number of involved items, this approach is computationally challenging, especially in the multidimensional case considered in this paper.

Despite their richness, TIMSS&PIRLS data are collected by a cross-section design, thus preventing studying the dynamics of the achievement process; see, among others, Kyriakides (2008). To overcome this limitation, several agencies, including INVALSI in Italy, are carrying out longitudinal surveys. The potentialities of longitudinal achievement data can be exploited by complex multilevel models, such as cross-classified multiple membership growth curve models [Grady and Beretvas (2010)] and multilevel latent Markov models [Bartolucci, Pennoni and Vittadini (2011)].

REFERENCES

AMMERMUELLER, A. and PISCHKE, J. S. (2009). Peer effects in European primary schools: Evidence from the progress in international reading literacy study. *J. Labor. Econ*. **27** 315–348.

ASPAROUHOV, T. (2006). General multi-level modeling with sampling weights. *Comm. Statist. Theory Methods* **35** 439–460. MR2274063

BARTOLUCCI, F., PENNONI, F. and VITTADINI, G. (2011). Assessment of school performance through a multilevel latent Markov Rasch model. *J. Educ. Behav. Stat*. **36** 491–522.

BOUHLILA, D. S. and SELLAOUTI, F. (2013). Multiple imputation using chained equations for missing data in TIMSS: A case study. *Large Scale Assess. Educ*. **1** 1–33.

CHIU, M. M. and XIHUA, Z. (2008). Family and motivation effects on mathematics achievement: Analyses of students in 41 countries. *Learn. Instr.* **18** 321–336.

FOX, J.-P. and GLAS, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* **66** 271–288. MR1836937

FOY, P. (2013). TIMSS and PIRLS 2011 user guide for the fourth grade combined international database. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA. Available at http://timssandpirls.bc.edu/timsspirls2011/international-database.html.

FOY, P. and O'DWYER, L. M. (2013). Technical Appendix B. School effectiveness models and analyses. In *TIMSS and PIRLS* 2011 *Relationships Among Reading*, *Mathematics*, *and Science Achievement at the Fourth Grade-Implications for Early Learning* (M. O. Martin and V. S. Mullis, eds.). TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA. Available at http://timssandpirls.bc.edu/timsspirls2011/downloads/TP11_Relationship_Report.pdf.

GOLDSTEIN, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education*: *Principles*, *Policy & Practice* **11** 319–330.

GOLDSTEIN, H. (2011). *Multilevel Statistical Models*, 4th ed. Wiley, New York.

GOLDSTEIN, H., CARPENTER, J. R. and BROWNE, W. J. (2014). Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *J. Roy. Statist. Soc. Ser. A* **177** 553–564. MR3249673

GRADY, M. W. and BERETVAS, S. N. (2010). Incorporating student mobility in achievement growth modeling: A cross-classified multiple membership growth curve model. *Multivar. Behav. Res*. **45** 393–419.

GRILLI, L. and RAMPICHINI, C. (2015). Specification of random effects in multilevel models: A review. *Qual. Quant*. **49** 967–976.

HAMMOURI, H. A. M. (2004). Attitudinal and motivational variables related to mathematics achievement in Jordan: Findings from the third international mathematics and science study (TIMSS). *Educ. Res*. **46** 241–257.

HANUSHEK, E. A. and WOESSMANN, L. (2011). The economics of international differences in educational achievement. In *Handbook of the Economics of Education* (E. A. Hanushek, S. Machin and L. Woessmann, eds.) **3**. Elsevier, The Netherlands.

ISTITUTO TAGLIACARNE (2011). Reddito e occupazione nelle province Italiane dal 1861 ad oggi. Istituto Tagliacarne, Roma.

JERRIM, J. and MICKLEWRIGHT, J. (2014). Socio-economic gradients in children cognitive skills: Are cross-country comparisons robust to who reports family background? *Eur. Sociol. Rev.* **30** 766–781.

JOHNSON, M. S. and JENKINS, F. (2005). A Bayesian hierarchical model for large-scale educational surveys: An application to the international assessment of educational progress. ETS Research report RR-04-38. Educational Testing Service, Princeton, NJ.

JONCAS, M. and FOY, P. (2013). Sample design in TIMSS and PIRLS. In *Methods and Procedures in TIMSS and PIRLS* 2011 (M. O. Martin and I. V. S. Mullis, eds.). TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA. Available at http://timssandpirls.bc.edu/methods/pdf/TP_Sampling_Design.pdf.

KIRSCH, I., DE JONG, J., LAFONTAINE, D., MCQUEEN, J., MENDELOVITS, J. and MONSEUR, C. (2002). *Reading for Change. Performance and Engagement Across Countries. Results from Pisa* 2000. OECD, Paris.

KREINER, S. and CHRISTENSEN, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika* **79** 210–231. MR3255117

KREUTER, F., ECKMAN, S., MAAZ, K. and WATERMANN, R. (2010). Children's reports of parents'education level: Does it matter whom you ask and what you ask about. *Surv. Res. Meth.* **4** 127–138.

KYRIAKIDES, L. (2008). Testing the validity of the comprehensive model of educational effectiveness: A step towards the development of a dynamic model of effectiveness. *Sch. Eff. Sch. Improv.* **19** 429–446.

LADD, H. and WALSH, R. (2002). Implementing value-added measures of school effectiveness: Getting the incentives right. *Econ. Educ. Rev.* **21** 1–17.

LI, K. H., MENG, X.-L., RAGHUNATHAN, T. E. and RUBIN, D. B. (1991). Significance levels from repeated p-values with multiply-imputed data. *Statist. Sinica* **1** 65–92. MR1101316

MARTIN, M. O. and MULLIS, I. V. S. (2012). Methods and procedures in TIMSS and PIRLS 2011. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA.

MARTIN, M. O. and MULLIS, I. V. S. (2013). Timss and Pirls 2011: Relationships Among Reading, Mathematics, and Science Achievement at the Fourth Grade-Implications for Early Learning. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA.

MISLEVY, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika* **56** 177–196.

RABE-HESKETH, S. and SKRONDAL, A. (2006). Multilevel modelling of complex survey data. *J. Roy. Statist. Soc. Ser. A* **169** 805–827. MR2291345

RAUDENBUSH, S. W. and WILLMS, J. D. (1995). The estimation of school effects. *J. Educ. Behav. Stat.* **20** 307–335.

REEVE, J. and JANG, H. (2006). What teachers say and do to support students' autonomy during a learning activity. *J. Educ. Psychol.* **98** 209–218.

RUBIN, D. (2002). *Multiple Imputation for Nonresponse in Sample Surveys*. Wiley, New York.

RUTKOWSKI, L., VON DAVIER, M. and RUTKOWSKI, D. (2014). *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*. Chapnam & Hall, Boca Raton.

RUTKOWSKI, L., GONZALEZ, E., JONCAS, M. and VON DAVIER, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educ. Res.* **39** 142–151.

SANI, C. and GRILLI, L. (2011). Differential variability of test scores among schools: A multilevel analysis of the fifth-grade Invalsi test using heteroscedastic random effects. *J. Appl. Quant. Meth.* **6** 88–99.

SCHAFER, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Stat. Neerl.* **57** 19–35. MR2055519

SNIJDERS, T. A. B. and BOSKER, R. J. (2012). *Multilevel Analysis*: *An Introduction to Basic and Advanced Multilevel Modeling*, 2nd ed. Sage Publications, Los Angeles, CA. MR3137621

STATACORP (2013). Stata: Release 13. Statistical Software. StataCorp LP, College Station, TX.

STONGE, J. H., WARD, T. J. and GRANT, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *J. Teach. Educ.* **62** 339–355.

TEKWE, C., CARTER, R., MA, C., ALGINA, J., LUCAS, M., ROTH, J., ARIET, M., FISHER, T. and RESNICK, M. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *J. Educ. Behav. Stat.* **29** 11–36.

TRANMER, M. and STEEL, D. G. (2001). Ignoring a level in a multilevel model: Evidence from UK census data. *Environ. Plann. A* **33** 941–948.

VON DAVIER, M., GONZALEZ, E. and MISLEVY, R. (2009). What are plausible values and why are they useful? In *ERI Monograph Series*: *Issues and Methodologies in Large-Scale Assessments* (M. von Davier and D. Hastedt, eds.) **2** 9–36.

WANG, Z., OSTERLIND, S. and BERGIN, D. (2012). Building mathematics achievement models in four countries using TIMSS 2003. *Int. J. Sci. Math. Educ.* **10** 1215–1242.

WEIRICH, S., HAAG, N., HECHT, M., BÖHME, K., SIEGLE, T. and LÜDTKE, O. (2014). Nested multiple imputation in large-scale assessments. *Large Scale Assess. Educ.* **2** 1–18.

WU, M. (2005). The role of plausible values in large-scale surveys. *Stud. Educ. Eval.* **31** 114–128.

YANG, M., GOLDSTEIN, H., BROWNE, W. and WOODHOUSE, G. (2002). Multivariate multilevel analyses of examination results. *J. Roy. Statist. Soc. Ser. A* **165** 137–153. MR1909740

L. GRILLI
C. RAMPICHINI
DEPARTMENT OF STATISTICS,
COMPUTER SCIENCE, APPLICATIONS "G. PARENTI"
UNIVERSITY OF FLORENCE
VIALE MORGAGNI 59
50134 FIRENZE
ITALY
E-MAIL: grilli@disia.unifi.it
rampichini@disia.unifi.it

F. PENNONI
DEPARTMENT OF STATISTICS
AND QUANTITATIVE METHODS
UNIVERSITY OF MILANO-BICOCCA
VIA BICOCCA DEGLI ARCIMBOLDI 8
20126 MILANO
ITALY
E-MAIL: fulvia.pennoni@unimib.it

I. ROMEO
LABORATORY OF ENVIRONMENTAL CHEMISTRY AND TOXICOLOGY
MARIO NEGRI
VIA LA MASA 19
20156 MILANO
ITALY
E-MAIL: isabella.romeo@marionegri.it