Dipartimento di
Scienze Economiche, Aziendali,
Matematiche e Statistiche "Bruno de Finetti"

Research Paper Series, N. 5, 2024

High School Proficiency of Future University Students: an Analysis based on INVALSI Data

Francesco Santelli[1], Gioia Di Credico[1], Claudia Di Caterina[2]

[1] Department of Economics, Business, Mathematics and Statistics, Università degli studi di Trieste
[2] Department of Economics, Università degli studi di Verona

email: gioia.dicredico@deams.units.it

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Research Paper Series

Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche “Bruno de Finetti”
Piazzale Europa, 1 – 34127 Trieste
Tel.: +390405587927
Fax: +390405587033
http://www.deams.units.it

EUT Edizioni Università di Trieste
Via E. Weiss, 21 – 34128 Trieste
Tel.: +390405586183
Fax: +390405586185
http://eut.units.it
eut@units.it

ISBN: 978-88-5511-535-3

High School Proficiency of Future University Students: an Analysis based on INVALSI Data

Francesco Santelli[1], Gioia Di Credico[1], Claudia Di Caterina[2]

[1]Department of Economics, Business, Mathematics and Statistics, Università degli studi di Trieste.
[2]Department of Economics, Università degli studi di Verona.

Contributing authors: fsantelli@units.it; gioia.dicredico@deams.units.it; claudia.dicaterina@univr.it;

Abstract

Large-scale assessment in the education field is key in every Country. In Italy, the institute that is in charge of evaluating pupils' proficiency is the INVALSI, via a set of standardized tests, that go in parallel with traditional school evaluation. Data collected in such a way at the individual level pose a statistical challenge, given the nested structure of students-classroom-school and the repeated measure longitudinal observations that are obtained for each student who performs the tests in several years during their scholastic career. We propose in this context the streamlined version of the mean-field variational Bayes (MFVB) algorithm for linear mixed models with crossed random effects in order to obtain plausible predictors of pupils' performances. The results and interpretation of model coefficients are in line with the literature on educational data.

Keywords: crossed design, education, scholastic assessment, mean field variational Bayes, random effects

1 Introduction

Evaluation of the Italian educational system's performance is carried out by a set of standardized procedures. For what concerns the school system, the evaluation is performed from a quantitative internal benchmarking perspective [1]. The definition of such analysis refers to the fact that benchmarking has also been applied to contexts that are supposed to be apart from the logic of profit. For that reason, it has been

adopted in private companies and public sectors [2].
The self-assessment concept is the core of internal quantitative benchmarking in the public sector. The idea is that the public sector can manage and produce by itself a considerable amount of data regarding its sub-entities to evaluate their performances. This benchmarking procedure often takes the name of *large scale scholastic assessment*, given the number of units involved and the fact that a set of standardized tools is used simultaneously over a considerable number of pupils.
Such assessment is carried out within Countries, to understand the health status of the educational system and to get an idea of the internal gaps, but also to compare Countries among them, and thus a set of comparable tests is delivered and analyzed. Testing different education systems in very diverse contexts is a massive challenge. The main institute in charge of that is PISA ("Program for International Student Assessment") inside the OECD framework [1] [3].
Scholars' main focus is, of course, on the proficiency test results, in many cases comparing regions, schools, subjects, and Countries. In addition, among others, some of the most common research questions deepened by researchers are related instead to such large assessment procedures and their consequences: the impact of evaluations on the way pupils approach the subjects and the way they learn/study, how teachers potentially change their behavior according to the tests, what the policies involved are, and the consequences of this whole process [4]. This means, in other words, that there are many implications involved, both from internal (within scholastic-education systems) and external (politics, universities, economy) perspectives [5]. It goes far beyond the mere study of pupils' proficiency. Many National Institutes that are linked to the PISA Organization, perform also other tests with national relevancy at different ages and stages of the scholastic career, and could also link such results with other data sources.
The Italian National Institute for the Evaluation of the School System *INVALSI* is in charge of such assessment in Italy: indeed, it gathers data from various sources and provides comprehensive analyses and reports, but of course, the main focus is to conceive, organize, and administer the proficiency tests to the pupils. Several publications and technical reports are available on the INVALSI website. Scholars have deeply explored data produced by INVALSI, for example, in the investigations on the gender gap in mathematics [6], [7], or the impact of the pandemic on the large-scale assessment governance [8]. Starting from INVALSI individual micro-data, several attempts have been made to produce statistical tools to properly analyze the proficiency values of students at different times and in different domains.
Such kinds of tests have been analyzed using a plethora of statistical techniques, depending on the data types and research aims. Most of the time, the inner hierarchical nature of the data (pupils are nested in classes, that are nested in schools, that are nested regions, and so on) poses a challenge to the statistical analyses, and models of the family *multilevel*, often called *random effects statistical models*, are proposed to deal with that. To this aim, we propose instead a Bayesian approach based on a streamlined mean field variational Bayes (MFVB) algorithm (see Section 3). We fit a

[1] PISA is the OECD's Programme for International Student Assessment. PISA measures 15-year-olds' ability to use their reading, mathematics, and science knowledge and skills to meet real-life challenges. More information is available at: OECD-PISA tests aims

crossed random effects model to account for two data variability levels: students and tests ([9]). Bayesian approaches are not entirely new to this kind of data, i.e. [10], but this specific model is first time introduced in the education field, being able to handle double variability levels and the estimation of precise credibility interval.

2 Dataset

The data drawn from the Italian "Anagrafe Nazionale della Formazione Superiore" have been processed according to the research project "From high school to the job market: analysis of the university careers and the university North-South mobility" carried out by the University of Palermo (head of the research program), the Italian "Ministero dell'Università e della Ricerca" (MUR), and INVALSI. This dataset, known as MOBYSU, has been collected to link university students' career to their scholastic proficiency. Several domain skills are tested on each pupil, such as Math, English language, Italian language, and Science.

Here, we focus on data from the cohort of pupils who graduated from high school in 2018/19 in Italy and then enrolled at an Italian university in the academic year 2019/20. Overall, such students are more than 240000. We select only those who attended for the first time the Italian school system and have no missing relevant information. A random sample of 14322 students, corresponding to 10% of the selected subset of interest, is then used for our analyses.

The response variable of our model refers to students' marks: two recorded at the end of the first school term (Italian and Math, oral) and two through the standardized INVALSI tests (Italian and Math, written), during their 10th and 13th high school grades (1). They all show a symmetrical behavior with a larger variability on the Math tests. For the two high school grades, Italian marks for the I term and Math marks for the INVALSI test are higher, on average.

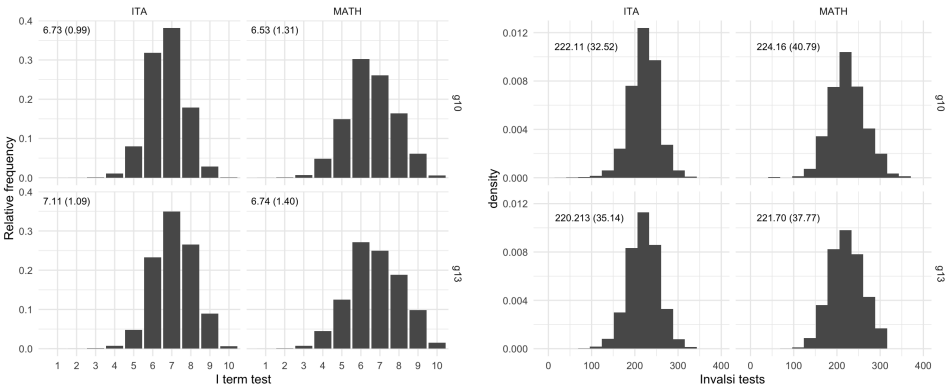

Fig. 1 Response variable distributions: Italian and Math first term evaluation marks (left) and Italian and Math INVALSI tests (right) for 10th (first row) and 13th (second row) high school grades. Annotations report the mean and the standard deviation (between brackets) of each variable.

Predictors involved in the analysis are listed in Tab.1. In detail, females constitute 57% of the observations, and only 3% of the students have non-Italian nationalities. Most students had a regular scholastic career (84%), while 11% of them started school one year before, and only 5% were held back one year. A vast majority of pupils (75%) attended Lyceum type of schools (40% scientific-technological; 12% linguistic; 11% classical; 12% other). Vocational institutions were instead chosen by 21% of the students. Around 4% attended private schools, while the remaining were public. Following the INVALSI classification of territorial units, 40% of the students are from Northern Italy (28% North-West; 20% North-East), 21% from the Center, and the others from South and South and Islands. Such classification is based on previous literature, studies, and state of art of the socio-economic divide, assigning Italian regions to five macro-areas that are different from the usual used by official Institutions such as Eurostat, and its classification of NUTS type. Regions are divided in the following way; North-East: Friuli Venezia Giulia, Trentino Alto-Adige, Veneto, Emilia Romagna. North-West: Lombardia, Piedmont, Liguria, Aosta Valley. Center: Lazio, Tuscany, Umbria, Marche. South: Campania, Abruzzo, Molise, Apulia. South and Islands: Basilicata, Calabria, Sicily, Sardinia.

The Economic, Social, and Cultural Status index (ESCS) gives information on students' family background with a mean value of 0.35 and a standard deviation of 0.94 at the individual level and 0.23 and 0.40 at the school level. Regarding parents' job positions, mothers are mainly unoccupied/retired (30%) or teachers/employees (27%). Fathers are less likely to be unoccupied/retired (5%), but they are involved in upper-level roles such as managers, university professors, and freelance positions (29%). We also ensured that the selected sample reflects the composition of the original population in terms of subgroups defined by key covariates, like geographic location and type of school.
Sample data in table 1 refer to the actual data used in model estimation (14322), while Complete data in table 1 are the initial data, more than 240000.

3 Methods and analysis

For each ith student, we assume the scores $\boldsymbol{y}_{ii'}$ on test i' follows a linear mixed model with two crossed random effects:

$$\begin{aligned} \boldsymbol{y}_{ii'}|\boldsymbol{\beta}, \boldsymbol{u}_i, \boldsymbol{u}'_{i'}, \sigma^2 &\overset{\text{ind.}}{\sim} N(\boldsymbol{X}_{ii'}\boldsymbol{\beta} + \boldsymbol{Z}_{ii'}\boldsymbol{u}_i + \boldsymbol{Z}'_{ii'}\boldsymbol{u}'_{i'}, \sigma^2\boldsymbol{I}), \quad i = 1, \dots, m, \\ \boldsymbol{u}_i|\boldsymbol{\Sigma} &\overset{\text{ind.}}{\sim} N(0, \boldsymbol{\Sigma}), \qquad \boldsymbol{u}'_{i'}|\boldsymbol{\Sigma}' \overset{\text{ind.}}{\sim} N(0, \boldsymbol{\Sigma}'), \quad i' = 1, \dots, m', \end{aligned} \tag{1}$$

where $\boldsymbol{X}_{ii'}$ is the $n_{ii'} \times p$ design matrix, $\boldsymbol{Z}_{ii'}$ and $\boldsymbol{Z}'_{ii'}$, respectively of dimension $n_{ii'} \times q$ and $n_{ii'} \times q'$, are the random effects matrices, $\boldsymbol{\beta}$ is the p-vector of fixed-effect coefficients, $\boldsymbol{u}_i$ and $\boldsymbol{u}'_{i'}$, respectively $q \times 1$ and $q' \times 1$, are the vectors of random effects, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$ are their $q \times q$ and $q' \times q'$ respective covariance matrices and σ^2 is the error variance.

The joint *a priori* density of the p fixed effects is $\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$. For the error variance σ^2 and the random effects covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$, we consider the

Table 1 Covariates introduced in the model estimation. Reference levels, in bold, are the first category of each variable. Percentages for categorical variables; range, mean (SD) for numerical variables.

Variable	Categories-Range	Sample data	Complete data
Gender	**Female**	57%	56%
	Male	43%	44%
Age	**Regular**	84%	84%
	Reception	11%	10%
	Failed	5%	6%
Nation	**Italian**	97%	93%
	Foreigner	3%	7%
School	**Public**	96%	97%
	Private	4%	3%
Student escs (EscsStud)	-3.91; +1.96	0.36 (0.94)	0.29 (0.96)
School escs (EscsSch)	-1.24; +1.81	0.23 (0.39)	0.18 (0.45)
School type (SchTy)	**Classic Lyceum**	13%	11%
	Scientific Lyceum	40%	35%
	Foreign Lang. Lyceum	12%	10%
	Others Lyceum	11%	15%
	Technical	3%	5%
	Vocational	21%	24%
Work Mother (Work.M)	**Unemployed**	30%	37%
	Worker	15%	11%
	Teacher/Employee	27%	29%
	Entrepreneur	11%	11%
	Manager/Self-Employed	17%	11%
Work Father (Work.F)	**Unemployed**	5%	2%
	Worker	22%	20%
	Teacher/Employee	18%	18%
	Entrepreneur	26%	25%
	Manager/Self-Employed	29%	36%
Year - treated as continuous	10^{th} **grade; -1.5**	50%	50%
	13^{th} grade; +1.5	50%	50%
Macro Area INVALSI	**Center**	21%	22%
	North-West	28%	26%
	North-East	20%	19%
	South	21%	23%
	Islands	10%	11%

following family of marginally non-informative prior distributions [11]:

$$\begin{aligned}
\sigma^2|a_{\sigma^2} &\sim \text{Inverse-}\chi^2(\nu_{\sigma^2}, 1/a_{\sigma^2}), \quad a_{\sigma^2} \sim \text{Inverse-}\chi^2(1, 1/(\nu_{\sigma^2} s^2_{\sigma^2})), \\
\boldsymbol{\Sigma}|\boldsymbol{A}_{\boldsymbol{\Sigma}} &\sim \text{Inverse-G-Wishart}(G_{\text{full}}, \nu_{\boldsymbol{\Sigma}} + 2q - 2, \boldsymbol{A}^{-1}_{\boldsymbol{\Sigma}}), \\
\boldsymbol{\Sigma}'|\boldsymbol{A}_{\boldsymbol{\Sigma}'} &\sim \text{Inverse-G-Wishart}(G_{\text{full}}, \nu_{\boldsymbol{\Sigma}'} + 2q' - 2, \boldsymbol{A}^{-1}_{\boldsymbol{\Sigma}'}), \\
\boldsymbol{A}_{\boldsymbol{\Sigma}} &\sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, \Lambda_{\boldsymbol{A}_{\boldsymbol{\Sigma}}}), \quad \boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}}} = \{\nu_{\boldsymbol{\Sigma}} \operatorname{diag}(s^2_{\boldsymbol{\Sigma},1}, s^2_{\boldsymbol{\Sigma},2})\}^{-1}, \\
\boldsymbol{A}_{\boldsymbol{\Sigma}'} &\sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, \Lambda_{\boldsymbol{A}_{\boldsymbol{\Sigma}'}}), \quad \boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}'}} = \left\{\nu_{\boldsymbol{\Sigma}'} \operatorname{diag}(s^2_{\boldsymbol{\Sigma}',1}, s^2_{\boldsymbol{\Sigma}',2})\right\}^{-1}.
\end{aligned} \tag{2}$$

In our application, the first group of random effects $\boldsymbol{u}_i$ $(i = 1, \ldots, m = 14322)$ corresponds to students enrolled at an Italian university in 2019/2020, and the second group $\boldsymbol{u}'_{i'}$ $(i' = 1, \ldots, m' = 4)$ corresponds to scores from assessments of Italian and Math skills. Specifically, for each student, an oral score was recorded at the end of the first term, and one written standardized INVALSI score was recorded at the end of the final term. Each combination student/test, corresponding to the pair (i, i'), is observed $n_{ii'} = n = 2$ times, namely in the 10th and 13th grades of high school. The design matrix $\boldsymbol{X}_{ii'}$ has $p = 26$ columns, including the intercept. Moreover, $q = q' = 2$ because we consider both random intercepts and random slopes for the two groups, meaning

$$\boldsymbol{Z}_{ii'} = \boldsymbol{Z}'_{ii'} = [1 \; x_{1,ii'j}]_{j=1,2}\,,$$

where $x_{1,ii'j} = 1, 2$ is the year indicator encoding the two high school grades. According to (1) and this set-up, the two scores of the ith student on the i'th test are modeled to be:

$$y_{ii'j}|\boldsymbol{\beta}, u_{0i}, u_{1i}, u'_{0i'}, u'_{1i'}, \sigma^2 \overset{\text{ind.}}{\sim} N(\beta_0 + u_{0i} + u_{0i'} + (\beta_1 + u_{1i} + u_{1i'})x_{1,ii'j} + \sum_{k=1}^{31} \beta_k x_{k,ii'j}, \sigma^2)$$

for $j = 1, 2$. The formula above shows that this modelling strategy allows for a different intercept and slope for every student/test combination. Heterogeneities among intercepts and slopes are defined by appropriate entries of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$.

Different product restrictions can be applied on the mean field approximation of the joint conditional density function of all parameters in (1) with covariance priors (3) ([12], Sect. 3):

$$q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}', \sigma^2, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}') = \begin{cases} q(\boldsymbol{\beta})q(\boldsymbol{u})q(\boldsymbol{u}')\, q(\sigma^2, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}'), & \text{restriction I,} \\ q(\boldsymbol{\beta}, \boldsymbol{u})q(\boldsymbol{u}')\, q(\sigma^2, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}'), & \text{restriction II,} \\ q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}')\, q(\sigma^2, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}'), & \text{restriction III.} \end{cases} \tag{3}$$

Product restriction I has the simplest streamlined implementation and scales well to very large problems, but may produce small posterior variances as it sets all posterior correlations between $\boldsymbol{\beta}$, $\boldsymbol{u}$ and $\boldsymbol{u}'$ to zero. Conversely, product restriction III allows for a full joint posterior covariance matrix of $(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}')$, leading to higher inferential accuracy but challenging computing that can be streamlined for limited m'. A

compromise is given by product restriction II, which includes posterior correlations between $\boldsymbol{\beta}$ and $\boldsymbol{u}$, for $\boldsymbol{u}$ larger than $\boldsymbol{u}'$.

For all the product restrictions, the specification of the prior distribution 3 leads to a full factorization of the q-densities related to the covariance matrix and auxiliary variables [12].

In the following, we focus on the product restriction III that allows for a full joint posterior covariance matrix of $(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}')$, leading to higher inferential accuracy but challenging computing that can be streamlined because the number of tests is small. Standard mean field variational Bayes steps (e.g. [13], Sect. 10.1–3) lead to the q-density functions of the model parameters having the following forms:

$$\begin{aligned} & q^*(\boldsymbol{\beta}, \boldsymbol{u}_{\text{all}}) \text{ has a } N\big(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})}\big) \text{ distribution,} \\ & q^*(\sigma^2) \text{ has an Inverse-}\chi^2\big(\xi_{q(\sigma^2)}, \lambda_{q(\sigma^2)}\big) \text{ distribution,} \\ & q^*(\boldsymbol{\Sigma}) \text{ has an Inverse-G-Wishart}(\mathrm{G}_{\text{full}}, \xi_{q_{(\boldsymbol{\Sigma})}}, \boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma})}) \text{ distribution} \\ \text{and} \quad & q^*(\boldsymbol{\Sigma}') \text{ has an Inverse-G-Wishart}(\mathrm{G}_{\text{full}}, \xi_{q_{(\boldsymbol{\Sigma}')}}, \boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma}')}) \text{ distribution,} \end{aligned} \tag{4}$$

where $\boldsymbol{u}_{\text{all}} = (\boldsymbol{u}, \boldsymbol{u}')$. The q-density parameters are obtained using a coordinate ascent algorithm running for 100 iterations.

INVALSI marks are standardized with respect to the national mean (200) and standard deviation (40). The national mean also corresponds to the central value of the INVALSI score sufficient class, indicating adequate skills and knowledge. This motivates us to center the first term scores to the sufficient mark (6) on the grade scale 0-10. This step permits obtaining a comparable range of the two types of test scores that also accommodate our prior distribution specification 3. We are aware that INVALSI standardized tests are conceived to be different from the internal evaluation of the school's teachers, but both are a proxy of pupils' proficiency and thus, if standardized properly, can be treated as different measures (using indeed different way of measuring proficiency) of the same topics and on the same subjects.

4 Main findings

Fixed marginal effects suggest that female students and those from Northern Italy regions perform the best (especially from the North-East), while South and South-Islands show the lowest proficiency, and this result is consistent with a plethora of studies in that respect (i.e., [14]). The socio-economic status dimension has a significant positive effect, as expected, especially at the individual level, but interestingly also at the school level. This means that a context effect influences individual performances, not only the individual social background. Each school, indeed, brings with it a lot of socio-economic characteristics of the place where it is located, and also its reputation plays a key role in attracting students belonging to a certain kind of social status. As highlighted in several papers, it can be seen as a sort of proxy of the overall context effects [15].

Lyceum pupils show the best scores on average, with Scientific Lyceum overperforming all the others. The effect of the parents' jobs depicts a pattern in which students with a father who is involved in teaching and a mother who is not working, maybe having more

Table 2 Random effects tests' estimates and standard deviations for tests and students (approximate posterior mean). Square root of diagonal entries of $\hat{\boldsymbol{\Sigma}}$ ($\hat{\boldsymbol{\Sigma}}'$) are denoted by $\hat{\sigma}$ ($\hat{\sigma}'$). The residual error standard deviation estimate is 0.848.

	Intercept	Slope
INVALSI - Ita ($\hat{u}'_1$)	-0.130	-0.057
INVALSI - Mate ($\hat{u}'_2$)	-0.084	-0.063
First term - Ita ($\hat{u}'_3$)	0.246	0.086
First term - Mate ($\hat{u}'_4$)	-0.032	0.033
$\hat{\sigma}'$	0.172	0.074
$\hat{\sigma}$	0.600	0.047

time to dedicate to the children, are the ones with the best results. On the contrary, when the mother is involved in demanding activities (entrepreneur, manager), pupils' scores tend to decrease. Overall, the effects related to the job position of the mother are way higher with respect to the ones related to the father. It highlights the crucial role played by the mother in the education-caring of the children; the more time mothers have, the more their children perform better. No clear effects are found for students one year ahead, but students who already failed at least one year are obtaining results way below average. Students who failed are not recovering, maybe lacking adequate support in order to do that. Pupils with no Italian nationality also perform below the mean, and in this case, the language barrier certainly plays a key role.
Regarding random effects, INVALSI items have a negative effect and first-term positive or close to zero effects on the intercept; thus, students, on average, are judged with more generosity by their own teachers, especially for what concerns Italian subjects, for which the difference between INVALSI evaluation and internal assessment is very large. On the other hand, for mathematics evaluations, the distance is not so big; in both cases, small but negative values show that mathematics is still the most difficult topic both as an internal assessment and as a standardized test. Slope variability is pretty low compared to intercept variability, so the effect is more on the starting level than on the overtime scores. But interestingly, internal evaluations have positive slopes, while INVALSI tests have negative slopes. There could also be an effect related to last year's exam in high schools, where the teachers could tend to judge with more generosity to give a chance to the pupils to sit for the final exam. Instead, standardized tests present an inverse behavior, showing an increasing level of difficulties in succeeding in the tests.

5 Conclusions

The work analyzed Italian students' proficiency data using the streamlined MFVB algorithm, exploiting the potentialities of such a method. The obtained results align with the literature on deepening proficiency data and, more specifically, Italian INVALSI data. Interestingly, given that it is known that results improve with the increase in socio-demographic status, some unexpected results emerged from the

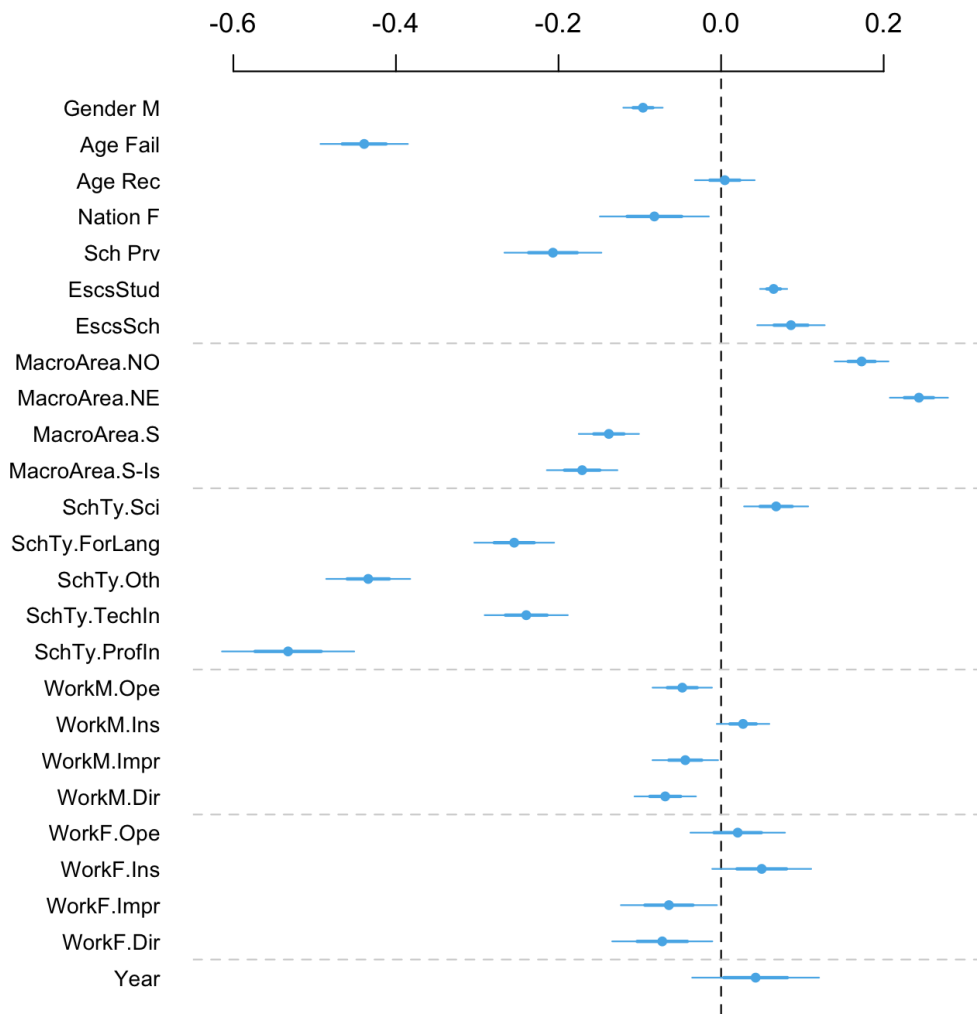

Fig. 2 Fixed effects: approximate posterior means (dots) and 95% credible intervals for the MFVB. The intercept estimate is equal to 0.809.

parents' jobs. It emerges that a different dimension deviates from the mere socio-demographic status, related to the awareness of the importance of having good marks (teachers have a positive effect) and more time to dedicate to children support (when a mother is unemployed, results improve). Some specific groups of students, such as recent immigrants with not yet Italian nationality or pupils that already failed at least one year, seem more vulnerable. The geographical South-North divide is clear from the findings, and the gender gap emerges, favoring females. This method allowed for fair comparisons among INVALSI performances and internal evaluations, highlighting different behavior in both intercept and slopes, with internal teacher being more

generous, especially at the last year of high school. Both assessments, anyway, showed how difficult is for Italian pupils to achieve good results in mathematics.

Declarations

Competing interests and funding

No competing interests and funding to declare.

Availability of data and materials

The code and data to reproduce the analysis are available on request.

Authors' contributions

F.S. and G.D.C. designed research; G.D.C. and C.D.C. analyzed data; F.S., G.D.C., and C.D.C. wrote the paper.

References

[1] Binder, M., Clegg, B., Egel-Hess, W.: Achieving internal process benchmarking: guidance from basf. Benchmarking: An Int. J. **13**, 662–687 (2006)

[2] Kouzmin, A., L"offler, E., Klages, H., Korac-Kakabadse, N.: Benchmarking and performance measurement in public sectors: Towards learning for agency effectiveness. Int. J. of Public Sect. Management **12**, 121–144 (1999)

[3] Sellar, S., Lingard, B.: The oecd and global governance in education. Testing Regimes, Accountabilities and Education Policy, 182–198 (2017)

[4] Wiseman, A.W., Taylor, C.S.: The Impact of the OECD on Education Worldwide. Emerald Group Publishing, ??? (2017)

[5] Niemann, D., Martens, K., Teltemann, J.: Pisa and its consequences: Shaping education policies through international comparisons. European Journal of Education **52**(2), 175–183 (2017)

[6] Cascella, C., Giberti, C., Bolondi, G.: An analysis of differential item functioning on invalsi tests, designed to explore gender gap in mathematical tasks. Stud. in Educ. Eval. **64** (2020) https://doi.org/10.1016/j.stueduc.2019.100819

[7] Giofré, D., Cornoldi, C., Martini, A., Toffalini, E.A.: population level analysis of the gender gap in mathematics: Results on over 13 million children using the invalsi dataset. Intell. **81** (2020) https://doi.org/10.1016/j.intell.2020.101467

[8] Milner, A.L., Mattei, P., Ydesen, C.: Governing education in times of crisis: State interventions and school accountabilities during the covid-19 pandemic. Eur. Educ. Res. J. **20**, 520–539 (2021)

[9] Baayen, R.H., Davidson, D.J., Bates, D.M.: Mixed-effects modeling with crossed random effects for subjects and items. J. Mem. Lang **59**(4), 390–412 (2008) https://doi.org/10.1016/j.jml.2007.12.005

[10] Trendtel, M., Robitzsch, A.: Modeling item position effects with a bayesian item response model applied to pisa 2009–2015 data. Psychological Test and Assessment Modeling **60**(2), 241–263 (2018)

[11] Huang, W.M.P. A.: Simple marginally noninformative prior distributions for covariance matrices. Bayesian Anal **8**, 439–452 (2013) https://doi.org/10.1214/13-BA815

[12] Menictas, M., Credico, D., G., W., P., M.: Streamlined variational inference for linear mixed models with crossed random effects. J. Comput. Graph. Stat. **32**, 99–115 (2022) https://doi.org/10.1080/10618600.2022.2096622

[13] Bishop, C.: Pattern recognition and machine learning. Springer google schola **2**, 35–42 (2006)

[14] Triventi, M., Argentin, G.: The north-south divide in school grading standards: New evidence from national assessments of the italian student population. Italian Journal of Sociology of Education **7**(Italian Journal of Sociology of Education 7/2), 157–185 (2015)

[15] Santelli, F., Ragozini, G., Vitale, M.P.: Assessing the effects of local contexts on the mobility choices of university students in campania region in italy. Genus **78**(1), 5 (2022)