Assessing maths learning gaps using Italian longitudinal data

Noname manuscript No.
(will be inserted by the editor)

Assessing maths learning gaps using Italian longitudinal data

Silvia Bianconcini · Stefania Mignani · Jacopo Mingozzi

Received: date / Accepted: date

Abstract In the educational context, one of the main goals is to reduce the disparities among students, generally at the national level, to allow all individuals to achieve a similar cultural background. Using data from a large-scale standardised test administered by INVALSI (National Institute for the Evaluation of the Educational System), this paper offers a first longitudinal analysis of the performance in the maths test of a cohort of students enrolled in 2013/2014 at grade 8 and observed up to grade 13. The aim is to identify those obstacles that undermine students' learning to help adopt informed educational actions. Specific features of these data are their hierarchical structure and the presence of not vertically scaled scores. Two approaches have been followed for their analysis: growth models and growth percentiles. Coherently with the literature, our results suggest the presence of a gender gap, a significant impact of the type of school, and of social-cultural background. Differently from previous research on the INVALSI data, we evaluate these time-invariant covariates' effects on students' performance over different school cycles.

Keywords INVALSI maths test · longitudinal data · student growth percentiles · gender gap · cross-cultural differences · school achievement

1 Introduction

With the assessment of students' skills, most advanced countries usually determine the quality and efficiency of their education system to identify critical

S. Bianconcini
Department of Statistical Sciences - University of Bologna
via delle Belle Arti, 41
40126 Bologna, Italy Tel.: +39-051-2098183
E-mail: silvia.bianconcini@unibo.it

S. Mignani · J. Mingozzi
Department of Statistical Sciences - University of Bologna

issues and adopt policies for improvement. One of the main goals is to reduce the disparities among students, generally at the national level, to allow all individuals to achieve a similar cultural background. Standardised evaluation systems that measure the skills and competencies acquired by the students permit to discover of whether there are inequalities between pupils within and/or between schools. Among the competencies needed to construct skills of the new generations, mathematical literacy is still a major challenge in basic education. It helps to produce functional and critical citizens [39,40], and promotes civic understanding and engagement in matters of crucial importance in modern societies [1,29,30]. All students should have the opportunity and the support to learn Mathematics with depth and understanding to open doors to productive futures. A lack of proper mathematical competencies keeps those doors closed. This issue also involves the critical aspect of gender disparities in mathematics achievement that have a remarkable impact on the enrolment into scientific degree courses and, consequently, on the job market [31,42].

Monitoring progress in the mathematical knowledge of the same generation of students provides insights to support different stakeholders in detecting and reducing knowledge gaps. Learning processes are cumulative, and policymakers have realised that single-year comparisons are inadequate to evaluate the school's ability to affect student academic progress. However, measuring the achievements of the same generation of pupils in different years of schooling is a long-term process.

Italy has adopted a national evaluation system managed by INVALSI (National Institute for the Evaluation of the Educational System). This Institute prepares, administers, and analyses the results of large-scale tests submitted to students at different grades to measure the knowledge acquired in fundamental skills. According to the current legislation, students of the 2nd, 5th, 8th, 10th, and 13th grades take tests on the Italian language (reading and grammar), Mathematics, and English language every year. Using the answers to the various test questions, INVALSI measures the outcome variable through a Rasch model. The cross-section analyses of these data provide only snapshots of the level of competencies acquired at a specific grade in a given year. A longitudinal approach could evaluate how the performances change over time. INVALSI make test results available at the various school grades. A unique code is associated with each student, so it is possible to trace the same cohort of pupils over time. Despite the issue's relevance, there is a lack of a comprehensive study in Italy that attempts to assess and identify the main determinants of the temporal acquisitions of mathematics competencies.

To the best of our knowledge, this paper provides a first longitudinal analysis of the performance in the INVALSI mathematics test of the cohort of students enrolled in 2013/2014 at grade 8, observed on three consecutive occasions, that is, up to grade 13. Previous dynamic analyses of the INVALSI data have considered repeated cross-section observations. The scores for different cohorts of students at different grades have been compared through autoregressive and difference-in-difference models [19,17,32].

Anyway, the INVALSI test scores are not vertically scaled at the different grades. To address this problem, two different approaches have been followed. The first is based on the estimation of a conditional three-level growth model that assesses the impact of covariates and accounts for the hierarchical structure of the data. The second approach consists of the estimation of student growth percentiles to study the normed growth of each student over time, accounting also for covariate effects. The joint use of these two approaches provides a more complete view of the temporal performance in the mathematics tests of the selected cohort of students and ensures greater robustness of the conclusions.

The results of both these models can be used for multiple purposes, but mainly from a policy-making viewpoint, to identify inequalities in the learning growths of different students. The analysis attempts to contribute to the current literature by detecting those obstacles that might undermine students' learning and thus can help adopt informed educational actions to reduce the negative impact of a specific student or school background. Our results corroborate some of the main findings already discussed in the literature, such as the gender gap, the school type's impact, and the role of the social-cultural background. However, from these previous researches, conclusions on the effects of these time-invariant covariates on the overall students' performance over different school cycles can also be drawn.

2 Dynamic analyses of educational data: main issues and previous researches on Italian data

The widespread availability of annual student assessment results during the last decade has greatly expanded the use of these data nationwide. Receiving particular interests are analyses of student academic growth [33,36,46].

In Italy, INVALSI associates a unique code (SIDI code) with each student, and this is used for all the tests at different grades. This allows us to follow the same cohort of students over different school cycles and use classical longitudinal models to properly investigate student growth over time. Nevertheless, none of the previous studies on the INVALSI test scores has used these longitudinal data.

In recent work, cross-sectional assessments at different schooling stages have been exploited to evaluate how achievement inequalities related to individual-ascribed characteristics develop over time [18]. The study analysed the Italian learning assessment of reading and math literacy carried out by the INVALSI on fifth and sixth graders in 2010 and 2011 on a sample of more than 30000 pupils. Gender, socioeconomic, immigrant background, and territorial inequalities in the transition between primary and lower secondary school were considered. Borrowing considerations from the econometric literature on repeated cross-sections [21,28,43], the authors provided a reflection on the advantages of pseudo-panel modelling for the study of the development of achievement inequalities as children progress into school and the conditions

for consistent estimation. The empirical analysis revealed that gender and socioeconomic inequalities widen in reading literacy but remain stable in maths, whereas the North-South inequality does not change in reading but severely increases in math. Immigrant background differentials are largely established at the end of primary school, with the exception of reading skills for boys. Second-generation immigrants do not lose ground with respect to natives in grade 8, and girls even catch up part of their disadvantage in math.

Continuing on the same line of research, the role of scaling issues entailed by using non-equated test scores at different stages of schooling has been explored to contribute to the literature on the correct use of international learning assessments in econometric modelling [17]. An in-depth analysis of difference-in-differences strategies have been conducted with the aim of evaluating the effect of institutional features on learning inequalities. The study showed that scaling issues might severely undermine the validity of the results delivered by difference-in-differences pooled individual-level models but do not apply to two-step estimation methods which represent a much better alternative.

Different dynamic analyses have been proposed for the study of the mathematical skills evolution in the INVALSI test of primary school pupils [32]. The cohorts involved were students born in 2004, who took the second test in 2012 and the fifth one in 2015, and students born in 2005, who took the second test in 2013 and the fifth test in 2016. The two cohorts were not strictly panel data since the students did not keep the same code from one test to another and there was no possibility to access their SIDI code. The dataset for each cohort was created by selecting the classes belonging to the national sample in both surveys. In the first step, the mathematics proficiency level of each student in the second and fifth grades was determined together with their difference. The possible relevance of the gender and geographical macro-area variables were then taken into account, but no significant differences in the distribution of the mathematics proficiency levels were detected. Overall, the mathematical ability assessed through the INVALSI test was considered to be in the process of being acquired gradually and changing over the course of school. To properly interpret the difference in maths skills between two successive grades, the scores must be longitudinally linked [34, 44], but INVALSI scores are not. To overcome this limitation, the second step of the analysis was focused only on items considered, for a series of criteria, sufficiently similar. This second-level analysis was used to help in identifying types of problematic situations for which there has been an increase or decrease in the ability of students in mathematics, in which school seemed to be adequately prepared or needed to be furtherly promoted.

Dynamic analyses of the INVALSI score have to account that it is neither vertically nor horizontally scaled. Vertical scales aim at measuring growth in student learning in terms of change in magnitude [41]. This is required by growth curve models, commonly applied in longitudinal data analysis [38, 15], where the outcome variable is modelled as a function of time. Several papers have examined the importance of the test score metric when individual student growth is the focus. Three-level hierarchical growth models have been applied

to vertically and not vertically scaled scores to compare accountability results [23]. It has been proved that the metric does not change the substantive inferences from ranking schools based on their fitted growth estimates but matters when the inferences pertain to actual absolute growth. Evidence also indicated that the metric does not affect evaluation results based on growth estimates that compare schools [37, 23].
The difficulty in interpreting scale scores to quantify growth has strongly hinted at trying to find a different metric to quantify change. Student growth percentiles [6] have become prominent in the analysis of student performances over time, especially in the US where the No Child Left Behind (NCLB) policy was adopted in 2002. In this framework, growth norms anchor understanding, using this familiar percentile metric to motivate discussions about whether student progress is normal or abnormal. These models provide an interpretation of growth in relative terms by comparing the student's growth with that of other students, which is, in essence, a normative interpretation [7]. Vertically-linked scales are no needed in this framework to make probabilistic comparisons about progress, providing information lacking in a simple analysis of scale score growths [15, 8].

Based on these findings, a three-level growth model is fitted to the INVALSI longitudinal data with the aim of drawing conclusions on the effect that student- and school-specific covariates can have on student achievements over time. The results cannot be used to derive conclusions on the individual growth pattern. To get insights into the student growth over time in mathematics competencies, growth percentiles are then estimated. In the latter, the individual student performance over time is compared with that of other students with similar score histories, known as academic peers. The rate of change is expressed in percentile, a commonly understood manner of comparing things to one another.

3 The dataset, the variables and some descriptive statistics

3.1 The dataset

The dataset used in the analyses was obtained from the population of students who took the mathematics tests in the school years 2013/2014, 2015/2016, and 2018/2019. As previously recalled, starting from 2011/2012, using the unique identification code (SIDI), it is possible to trace each student's results in the national tests taken at different grades. The selected cohort of students was observed from the last year of the first education cycle and throughout the second cycle. Indeed, beginning in 2013/2014, the INVALSI mathematics test was also administered at the final state exam of the first education cycle. Then, the three consecutive occasions considered in this study refer to 2013/2014, which represents the last year of the first cycle of education for this cohort of students (grade 8), 2015/2016 is the second class of the upper secondary school (grade 10), and 2018/2019 that, for this cohort, represents the last year

of the second cycle of education (grade 13).

A previous data cleaning has been performed. Through a listwise deletion, only the students with all the information available on each occasion are included in the data set. After this procedure, the cohort was reduced to 34544 students of 2808 secondary schools.

3.2 The variables

The aim is to study the influence of several factors at the individual and school level that are known to affect school achievements. The response variable measuring the math skills acquisition is represented by the Rasch maths score that INVALSI systematically computes for each pupil every time they take the national test. This score is standardised to be on average equal to 200, with a standard deviation of 40, at the national level. This means that, for each individual, values greater than 200 indicate student maths skills above the national average, whereas lower values indicate worse performance.
The variables used to discriminate the performances in the analyzed cohort are available at both the student and school levels. For the former, the information relates to the gender (male, female *ref.*), the immigration status (native *ref.* - that are Italian students, first and second-generation foreigners) and the Economic Social Cultural Status (ESCS) of each student. Theoretical considerations have mainly driven the choice of these control variables. Several studies have confirmed gender differences in school performance in Italy (see, among others, [10,22,24,18]). Boys tend to outperform girls in maths and science, whereas female students tend to perform better than males in Italian language [25,27]. The educational integration of immigrant children is a fundamental step for the economic integration of the immigrants in the host societies [35]. However, Italy is characterised by essential disparities between foreign and native students that also persist after accounting for family background characteristics as well as unobserved heterogeneity [4]. Among the family characteristics known to affect student achievements, socioeconomic and cultural status, as measured by the ESCS index, plays a prominent role [16]. As computed by INVALSI, this indicator is standardised to have a zero mean and unit standard deviation at the national level. Higher and positive values indicate a higher socioeconomic status.

We have also accounted for several school features, such as the type of school (Scientific Lyceums *ref.* - Other Lyceums, together since they share the same level of mathematical competencies - Technical Institutes - Vocational Institutes), the geographical macro-area in which it is located (North East (province of Bolzano, province of Trento, Veneto, Friuli-Venezia Giulia, Emilia Romagna) - North West (Vallle d'Aoasta, Piemonte, Lombardia, Liguria) - Center (Toscana, Umbria, Marche, Lazio) *ref.* - South (Abruzzo, Molise, Campania, Puglia) - South and Islands (Basilicata, Calabria, Sicilia, Sardegna)), and the associated ESCS index. As detailed by many studies [**?**,10, 3,2], the Italian education system is characterised by an evident North-South

gap, with students attending schools in North and Central Italy performing better than those in the Southern regions. The type of school is another determinant of student achievements [10].

3.3 Some descriptive statistics

Preliminary analyses are helpful in underlining the key characteristics of the selected cohort of students.

Table 1: Descriptive statistics of the (time-invariant) characteristics for the observed cohort of students

Variable	Description	
Explanatory variables - student level		
Male	Percentage males	0.47
Female	Percentage females	0.53
Native	Native (percentage)	0.93
2G	Second-generation migrant (percentage)	0.04
1G	First-generation migrant (percentage)	0.03
ESCS	Economic Socio-Cultural Status Index	0.10
	(average over the time span - st. dev. in brackets)	(0.95)
Explanatory variables - school level		
ScLy	Scientific Lyceums (percentage)	0.26
OtLy	Other Lyceums (percentage)	0.28
TeIn	Technical Institutes (percentage)	0.31
VoIn	Vocational Institutes (percentage)	0.14
North-West	Located in North-West Italy (percentage)	0.21
North-East	Located in North-East Italy (percentage)	0.23
Center	Located in Central Italy (percentage)	0.21
South	Located in South Italy (percentage)	0.22
South and Islands	Located in South and Islands (percentage)	0.14
ESCS school	Economic Socio-Cultural Status Index	0.09
	(average over the time span - st. dev. in brackets)	(0.41)
Number of units		34544
Number of schools		2808

Table 1 reports descriptive statistics for the time-invariant features. The cohort is slightly unbalanced in favour of females (53%), and the great majority of students are native pupils (93%). The highest proportion of students is enrolled in Technical institutes (31%), whereas only 14% of students attend Vocational schools and one in four students is enrolled in schools with advanced mathematics curricula (Scientific lyceums). The distribution of students by geographical area shows similar percentages (around 20-22%) for all the areas except for South and Islands (14%). The socioeconomic background shows more variability among students than among schools, confirming the role of socioeconomic characteristics in conditioning school choice.

Table 2 contains the average total scores in the different years and their average values by different groups of students and schools. Standard deviations are reported in brackets. The average maths score shows a slight difference in

Table 2: Average mathematics score by different groups of students and schools (standard deviation in brackets).

Variable	2013/2014	2015/2016	2018/2019
maths score	212.02 (38.53)	210.13 (41.23)	209.73 (39.55)
Gender			
Male	217.35 (39.70)	217.19 (42.92)	216.39 (41.00)
Female	204.29 (36.82)	203.86 (38.51)	203.81 (37.24)
Immigration status			
Native	212.83 (38.51)	210.78 (41.27)	210.20 (39.64)
2G	204.17 (37.45)	204.47 (39.81)	206.80 (38.24)
1G	195.80 (35.59)	196.25 (38.35)	198.14 (36.40)
Type of school			
ScLy	-	243.78 (38.29)	239.55 (35.13)
OtLy	-	199.46 (34.68)	200.87 (32.82)
TeIn	-	26.25 (34.99)	208.29 (35.10)
VoIn	-	177.79 (28.05)	175.50 (29.81)
Macro-area			
North-West	-	219.68 (39.85)	221.58 (36.96)
North-East	-	220.02 (39.32)	222.59 (39.71)
Center	-	207.88 (41.15)	207.54 (39.19)
South	-	201.91 (40.49)	198.70 (38.41)
South and Islands	-	196.24 (40.19)	191.95 (37.46)

the three grades even if it always remains above the national mean, set to 200, with a lower value for grade 13. In terms of variability, on each occasion, the standard score deviation is close to 40, which is the value of the standard deviation at the national level. Males have a better performance on average than female students on each occasion. This difference persists over time with a larger gap in the last observed year. The geographical area also plays an important role in explaining the potential variability existing between schools located in different areas. Specifically, there is a clear worse performance in maths in the Southern schools with respect to the Northern ones. In the school year 2015/2016, the average test score has always been above the national average, while in grade 13 the students from schools located in Southern Italy have shown, on average, lower mathematical knowledge. Important differences are also observed in students attending different secondary schools. Students in Scientific lyceums have higher average scores than in other schools, with more relevant differences observed for students attending Vocational institutes.

4 Growth models for educational longitudinal analyses

Various methods have been developed to assess student achievements over time, including value-added analyses, growth-to-standard analyses, growth curve modelling, value tables, gain scores, and student growth percentiles.

The variety of models is strictly connected with the various questions and purposes that longitudinal analyses are designed to address [15].

Educational data observed over time have a natural three-level hierarchical structure: occasions nested into students and nested into schools. This requires accurate methodologies to account for the dependence structure in these data and multilevel growth models for nested longitudinal data are a natural choice [38,20]. In these models, the outcome variable is required to be vertically scaled to provide a cross-grade achievement continuum that enables comparisons of student performances on different occasions. Nonvertical scales do not allow for such comparisons [11,12,13,14]. INVALSI scores are neither horizontally nor vertically scaled, and this should prevent the application of hierarchical growth models to these data.

Several papers have examined the impact of using theoretically non-optimal assessments and metrics as a basis for monitoring school performance using growth models [23]. Specifically, these studies have analysed the effect on results when inferences are based on not vertically scaled scores. Overall, inferences about individual students' growth will be biased, but for school accountability or evaluation purposes, results are consistent across metrics and assessments, with metric effects being negligible and assessment effects being minimal to moderate, depending on the model. Using longitudinal models for school accountability gives researchers and policymakers more flexibility since inferences based on non-scaled scores provide statistically and substantively similar results to inferences based on scaled scores.

Based on these findings, a three-level growth model is fitted to the INVALSI longitudinal data with the aim of drawing conclusions on the effect that student- and school-specific covariates can have on student achievements over time. The results cannot be used to derive conclusions on the individual growth pattern. To get insights into the student growth over time in mathematics competencies, student growth percentiles are then estimated. In the latter, the individual student performance over time is compared with that of other students with similar score histories, known as *academic peers*. The rate of change is expressed in percentile, a commonly understood manner of comparing things.

4.1 Multilevel growth models for nested longitudinal data

Let y_{ijt} be the maths score achieved at time t by the jth student in the ith school, with $t = 8$, 10, and 13 grade, $j = 1, \ldots, 34544$, and $i = 1, \ldots, 2808$. With only three occasions available, the student performance in mathematics over time is assumed to follow a linear trajectory. That is,

$$y_{ijt} = \pi_{0ij} + \pi_{1ij} \cdot (t - 8) + \varepsilon_{ijt} \tag{1}$$

where π_{0ij} and π_{1ij} are the intercept and slope of the growth trajectory of the j-th student attending the i-th school. The error term ε_{ijt} is assumed to follow a normal density with zero mean and constant variance σ^2_ε, being uncorrelated

over time and between different students. That is, $COV(\varepsilon_{ijt}, \varepsilon_{ij't'}) = 0, \forall j \neq j'$, and $i \neq i'$. The variability of the individual trajectories both at the initial status and in the growth rate is described as follows

$$\pi_{0ij} = \beta_{00i} + \boldsymbol{\beta}_{01}\mathbf{w}_j + r_{0ij} \tag{2}$$
$$\pi_{1ij} = \beta_{10i} + \boldsymbol{\beta}_{11}\mathbf{w}_j + r_{1ij}. \tag{3}$$

β_{00i} and β_{10i} represent the intercept and slope of the specific linear trajectory of the school i, whereas $\mathbf{w}_j$ is the three-dimensional vector of the individual characteristics observed in this dataset, that is gender, immigration status, and ESCS index. The corresponding vectors of coefficients, $\boldsymbol{\beta}_{01}$ and $\boldsymbol{\beta}_{11}$, represent the difference in the expected initial status and rate of change, respectively, in different groups of students identified by combinations of these individual covariates. The error components r_{0ij} and r_{1ij} quantify the differences in the intercept and slope of each individual with respect to the specific average pattern of the group at which he/she belongs. They are assumed to follow a bivariate normal density with zero mean vector and full covariance matrix $\begin{bmatrix} \tau_{\pi_{00}} & \tau_{\pi_{01}} \\ \tau_{\pi_{10}} & \tau_{\pi_{11}} \end{bmatrix}$. At the school level, we specify the following equations

$$\beta_{00i} = \gamma_{000} + \boldsymbol{\gamma}_{001}\mathbf{w}_i + u_{00i} \tag{4}$$
$$\beta_{10i} = \gamma_{100} + \boldsymbol{\gamma}_{101}\mathbf{w}_i + u_{10i}, \tag{5}$$

where γ_{000} and γ_{100} are the expected intercept and rate of growth of the trajectory defined for the overall population, that is common to all the students and schools. The three-dimensional vector $\mathbf{w}_i$ contains the time-invariant school covariates, that is the geographical macro-area, the type of school, and ESCS. The corresponding vectors of coefficients, $\boldsymbol{\gamma}_{001}$ and $\boldsymbol{\gamma}_{101}$, capture the expected shift in the intercepts and slopes of different groups of schools identified through combinations of the covariates. As for the second level, the errors u_{00i} and u_{10i} are assumed to be normally distributed with zero mean vector and full covariance matrix, denoted by $\begin{bmatrix} \tau_{\beta_{00}} & \tau_{\beta_{01}} \\ \tau_{\beta_{10}} & \tau_{\beta_{11}} \end{bmatrix}$.

The estimation of the model when the scores y_{ijt} are not vertically scaled prevents providing a direct interpretation of the growth parameters γ_{000} and γ_{100}, but it allows to get insights into the role that student- and school-specific covariates have on the acquisition of the mathematics competencies over time. The focus is then on the parameters $\boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{11}, \boldsymbol{\gamma}_{001}, \boldsymbol{\gamma}_{101}$ and their significant effect on the student growth trajectories.

4.2 Student growth percentiles

To draw conclusions on the growth in the mathematics competencies of the observed cohort of students, a different measure of growth should be considered. Student growth percentiles have been developed to provide a general measure, going beyond the common (mis)conception that to measure student growth in

education, the subject matter and grades over which growth is examined must be on the same vertical scale. Not only is a vertical scale always necessary, but its existence can obscure fundamental concepts necessary to understand growth. Vertically scaled scores are necessary to measure the magnitude of growth, but not growth in general [6,45].

The search for a description regarding the change in achievement over time is best served by considering a normative quantification of student growth. For each student, growth is quantified in comparison with that of students that share a similar score history, known as *academic peers*. Conditioning upon prior achievements defines a distribution of outcomes on the t-th grade test given specific values of the scores at the previous grades, that is first, second up to the $(t-1)$-th. This conditional distribution provides the context within which a student's achievement at grade t can be understood normatively. Students with achievement in the upper tail of this distribution have demonstrated high rates of growth relative to their academic peers, whereas those students with achievement in the lower tail of the distribution have demonstrated low rates of growth. Students with current achievement in the middle of the distribution could be described as demonstrating "typical" growth. Because past scores are used solely for conditioning purposes, one of the major advantages of using growth percentiles to measure change is that estimation does not require a vertical scale.

Calculation of a student's growth percentile is based upon the estimation of the conditional density associated with a student's score at time t, say y_{it}, using the student's prior scores at previous occasions, that is $y_{i1}, y_{i2}, \ldots, y_{i,t-1}$, as the conditioning variables. The student's growth percentile is then defined as the percentile of the score in this conditional density. The percentile reflects the likelihood of such an outcome given the student's prior achievements.
Quantile regressions are used to estimate features of the conditional distribution of the student [26]. In particular, one estimates the conditional quantiles for all possible test score histories, which are then used for assigning percentile ranks to students. In general, the τ-th conditional quantile is the value $Q_{y_{it}}(\tau \mid y_{i,t-1}, \ldots, y_{i1})$ such that

$$P[y_{it} \leq Q_{y_{it}}(\tau \mid y_{i,t-1}, \ldots, y_{i1})] = \tau. \tag{6}$$

The conditional quantiles are then modelled for achievement scores as follows

$$Q_{y_{it}}(\tau \mid y_{i,t-1}, \ldots, y_{i1}) = \sum_{j=1}^{t-1} \sum_{k=1}^{K} \phi_{ik}(y_{ij}) \beta_{ik}(\tau), \tag{7}$$

where ϕ_{ik} denotes B-spline basis functions of prior test scores. Following common tradition, bases consisting of seven cubic polynomials are used to smooth irregularities found in multivariate assessment data [8]. The B-spline functions are chosen to improve the model fit by adding flexibility in the treatment of prior test scores as covariates, primarily in that they allow for nonlinearities in the relationship between current and prior scores.

One hundred quantile regressions are estimated, one for each percentile. Regressions are run separately for each grade and year. Conditional test scores are estimated for each percentile by generating fitted values from the regressions as follows

$$\hat{Q}_{y_{it}}(\tau \mid y_{i,t-1}, \ldots, y_{i1}) = \sum_{j=1}^{t-1} \sum_{k=1}^{7} \phi_{ik}(y_{ij}) \hat{\beta}_{ik}(\tau) \tag{8}$$

A student's conditional percentile rank is then computed by counting the number of conditional percentiles that result in fitted test scores that are smaller than the student's current grade test score, y_{it}.

Once conditional percentile ranks are computed for all students, a score can be assigned to each school by computing the median or mean conditional percentile rank of the students within their classes. These scores can be used to form rankings of schools by their estimated effectiveness. An attractive feature of growth percentile models is that once computed, the student growth percentiles can be used to provide a variety of descriptive portraits [6].

5 Results

5.1 Three-level conditional growth model

The longitudinal analysis based on multilevel growth models allows us to discriminate the temporal performance of different groups of students. Critical characteristics that can help policymakers to adopt informed strategies to mitigate and reduce disparities in the student learning process can also be determined. All the subsequent analyses are performed using the `lme4` package of the R software [5].

Given the importance of accounting for the influence of both individual and school characteristics, a conditional three-level growth model has been estimated. As discussed before, due to the not vertical scaled nature of the INVALSI maths score, Table 3 reports the fixed effect estimates of the models since they are the only ones that can be properly interpreted and discussed.

There are significant differences in the eighth grade according to the gender, immigration status of the students, and their ESCS levels. However, the corresponding slopes are not significant. Male students have significantly better performance in the maths test in the eighth grade than females. Still, their rates of growth are not significantly different implying a constant gap in the performance of male and female students, defined by the discrepancy in the average maths scores observed in the eighth grade. These analyses would suggest that it is important to tackle the gender gap in the early stages, specifically to detect when the gap first appears. The same considerations can also be extended to immigration status, being the math performances of native students and foreigners significantly different only in the eighth grade.

Considering the school level covariates, there is an evident different performance among students attending different types of school. Students in the

Table 3: Fixed-effect estimates of the three-level conditional growth mode.

	Parameter	Estimate	St. err.	t
Intercept	γ_{000}	193.80	1.280	151.340***
Time	γ_{100}	1.428	0.237	6.022***
Gender	β_{01}	9.068	0.396	22.896***
1G	β_{02}	−6.349	1.082	−5.867***
2G	β_{03}	−5.172	0.980	−5.281***
ESCS student	β_{04}	1.158	0.208	7.592***
ESCS school	γ_{001}	8.251	0.983	8.391***
North West	γ_{002}	3.958	0.925	4.280***
North East	γ_{003}	3.808	0.948	4.019***
South	γ_{004}	0.152	0.951	0.160
South and Islands	γ_{005}	0.817	1.049	0.779
OtLy	γ_{006}	−20.01	0.663	−30.191***
TeIn	γ_{007}	−21.93	0.787	−27.882***
VoIn	γ_{008}	−35.79	1.027	−34.833***
Time: Gender	β_{11}	−0.032	0.071	−0.455
Time: 1G	β_{12}	0.493	0.193	2.558**
Time: 2G	β_{13}	0.299	0.175	1.713*
Time: ESCS student	β_{14}	−0.042	0.037	−1.126
Time: ESCS school	γ_{101}	−0.001	0.194	−0.005
Time: North West	γ_{102}	1.588	0.193	8.250***
Time: North East	γ_{103}	2.081	0.199	10.422***
Time: South	γ_{104}	−1.124	0.198	−5.671***
Time: South and Islands	γ_{105}	−2.146	0.219	−9.802***
Time: OtLy	γ_{106}	−3.749	0.123	−30.496***
Time: TeIn.	γ_{107}	−1.946	0.148	−13.181***
Time:VoIn.	γ_{108}	−4.001	0.192	−20.792***

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$

AIC: 986437.4, BIC: 986781.2, Log-likelihood: −493182.7, Deviance: 986365.4.

Scientific lyceums have higher performances on the first occasion than students attending the other three types of schools. Their trajectories have positive and higher slopes than the other schools. This implies that students with more solid mathematics knowledge tend to enrol in Scientific lyceum, whereas pupils with more modest mathematical abilities generally enrol in Technical institutes or Other lyceums. Finally, those who have little skills in Mathematics tend to enrol in Vocational institutes.

It is evident that, in the eighth grade, there are significant differences in the maths performances of students attending schools located in different macro-areas, except in the South and South - Islands. Students living in Southern regions have similar performances in the maths test in the eighth grade to students attending schools located in Central Italy. However, these Southern students have significantly different growth rates from those living in the Central regions. In particular, even if there is no difference in the eighth grade, Southern students have a smaller rate of growth with respect to those students attending schools in the Central regions.

5.2 Growth percentiles analyses

5.2.1 Student growth

To understand the academic progress of the selected cohort of students from the eighth to the thirteenth grade, a student growth percentile model is estimated. All the analyses are done using the `SGP` package of the R software [9].

This approach helps in understanding the heterogeneity in the performances of different groups and supports the identification of effective practices that could help students attain higher academic performances.

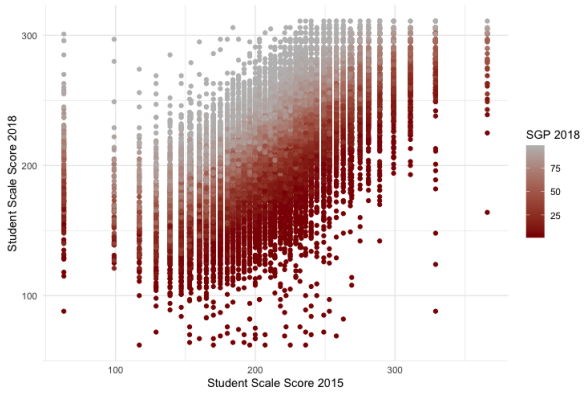

Fig. 1: Growth chart of the cohort of students at grade 13 compared with score obtained at grade 10.

Student growth percentiles provide a measure of student progress that compares changes in a student's INVALSI maths test scores at grade 13 to changes in the INVALSI maths test scores of other students with similar scores in previous grades. Percentiles are commonly understood values that express the percentage of cases that fall below a certain score. Figure 1 provides a graphical representation of the performance in the test score of the cohort of students from grades 10 to 13. The coloured shades represent the percentiles associated with each student. Given a specific value obtained in the INVALSI maths score in the tenth grade, it is possible to evaluate the percentile of a given student based on its score in grade 13. As an example, a student with a score in grade 10 equal to 300 will fall into the first quartile if his/her score in 2018 is around 200, whereas he/she will be located in third quartile if the thirteen grade score is around 300.

The student growth percentile is useful to better understand students' performances. There is a story behind every student growth percentile, and educators are encouraged to seek out these stories.

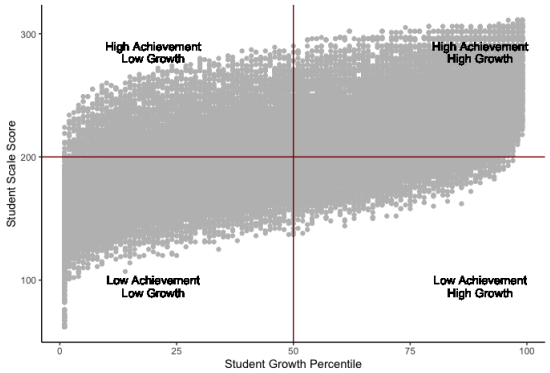

Fig. 2: Student performance on the INVALSI mathematics test at grade 13: student growth from grade 8 INVALSI test to grade 13 test (*x-axis*), and INVALSI mathematics test score at grade 13 (*y-axis*).

Figure 2 illustrates the performance in the mathematics test at grade 13 for the entire cohort of students, accounting for the growth in the mathematics test from grade 8 to grade 13. Each dot on the graph represents an individual student. The vertical axis represents student achievement on grade 13 INVALSI mathematics test, and the horizontal axis represents student growth from the grade 8 INVALSI mathematics test to the grade 13 test. Therefore, students shown in the upper right quadrant of the graph demonstrated higher growth and higher achievement than their peers with similar score histories. By contrast, students in the lower left quadrant demonstrated lower growth and lower achievements than their peers with similar INVALSI test score histories. The figure shows that among the students with low achievement in grade 13 most also have low growth. On the other hand, among students with high scores in grade 13, most have high growth. These results can be interpreted as no success in overcoming differences in subsequent grades for those students with a low level of achievement in grade 8.

To discriminate the growth behaviour of the cohort of students, the individual specific covariates can be taken into account. In Figure 3, the blue dots represent male students, while the red dots indicate female students. It is evident that male students mainly obtain scores above the national average in grade 13 and generally higher than those obtained by female students. However, in terms of growth, both groups of students are almost equally distributed on the left and right of the vertical line indicating the median percentile at the national level, which is 50. This is also confirmed in the bar chart reported in Figure 3 (*right*) which illustrates the proportion of students who grew in three clusters of growth percentile values distinguished with respect to student gender. Student growth is defined as "Low" if his/her percentile is smaller than 40. A "typical" growth is associated with a student growth percentile ranging

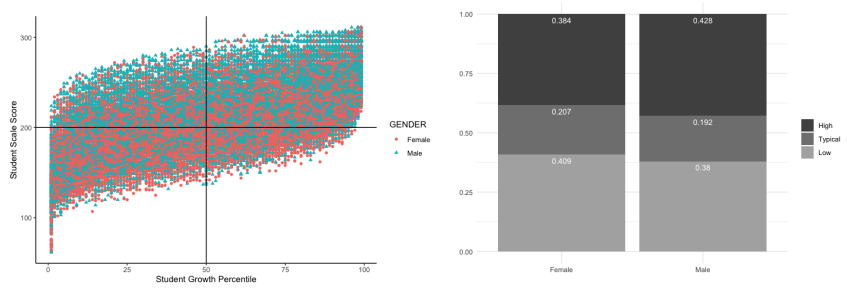

Fig. 3: Student performance (*left*) and growth distribution (*right*) by gender.

from 40 to 59, whereas an "high" growth refers to percentiles greater than or equal to 60. These distributions are very similar between males and females for typical and low growth. There is only a slight difference in proportion for high growth. .

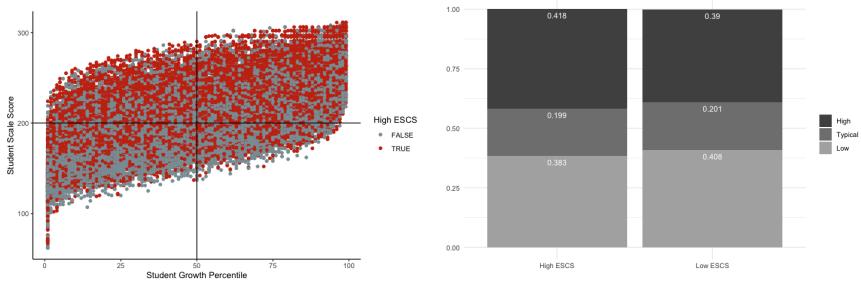

Fig. 4: Student performance (*left*) and growth distribution (*right*) by socioeconomic status (ESCS index).

Figure 4 discriminates the student performance by including the student's socioeconomic status. The red dots represent students with high socioeconomic status, with an ESCS index greater than the overall average ESCS, while the grey dots indicate students with low socioeconomic status. Students with a high socioeconomic status are generally characterised by higher INVALSI maths scores than students with lower socioeconomic status. However, as shown by the growth distribution in Figure 4 (*right*), there are no relevant differences in growth between the two groups of students. These findings seem to highlight that the school system is not able to take under control and reduce, over the entire formative period, the impact of the socioeconomic context.

These analyses have also been performed by accounting for the immigration status of the students. Conclusions are similar to the ones derived via the estimation of the three-level growth model, so they are not displayed for

space reasons. Better performances in terms of achievements are observed for the native students and worse for the first-generation foreigners. However, no relevant differences are highlighted in terms of growth.

5.2.2 School growth analyses

Student growth percentiles enable educators to chart the growth of an individual student compared to that of academic peers. They can also be aggregated to understand growth at the school level. To summarise student growth rates by schools, individual student growth percentiles can be aggregated. The most appropriate measure for reporting growth for a group is the median student growth percentile.

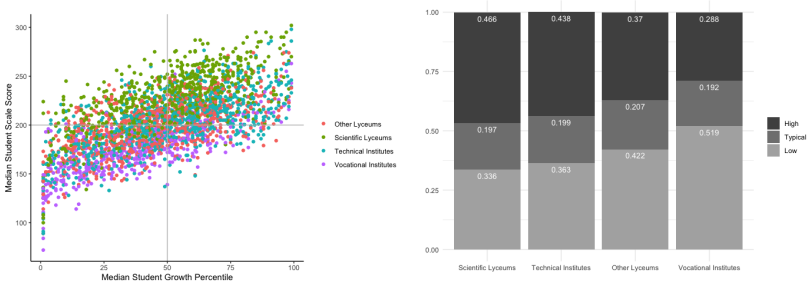

Fig. 5: School achievements (*left*) and growth (*right*) by type of school.

In Figure 5 (*left*), each dot represents one of the 2808 schools. School achievements, displayed on the vertical axis, are determined by the median of the INVALSI mathematics test scores obtained by students in that specific school. Growth, which is placed on the horizontal axis, is determined by finding the median student growth percentile in a school. The horizontal and vertical lines are placed at the national average score, equal to 200, and at the national median percentile, equal to 50, respectively. Figure 5 compares the performance of the different types of schools. As outlined in the three-level growth analysis, Scientific lyceums are characterised by high median INVALSI math scores but also by students that present the highest growth levels. Conversely, Vocational institutes have the worse performance being mainly located in the bottom left panel, with lower median scores and percentiles. In terms of growth distribution, Figure 5 (*right*) shows that median percentiles are higher in Scientific lyceum, followed by Technical institutes, Other lyceums, and Vocational institutes that have the highest percentage of median growth below the 39-th percentile.

Finally, Figure 6 shows achievements (*left*) and growth (*right*) for schools located in different geographical areas. The results corroborate the well-known differences between schools located in the North respect to those in Southern

Italy. The former have the highest median scores and highest median percentiles, in particular for schools situated in the Northern East. On the other hand, schools in Southern Italy and in the Islands are characterised by median scores that are lower than the average national score (equal to 200) and a median percentile smaller than 50. Differences in the growth distribution are highlighted in Figure 6 (*right*), where schools located in Northern Italy are characterised by a high level of growth whereas for those in the South there is a high percentage (around 50%) of schools with a median percentile smaller than 40.

The territorial gap is one of the main problems of the Italian educational system, and specific educational actions should be supported to reduce disparities, especially in Vocational schools.

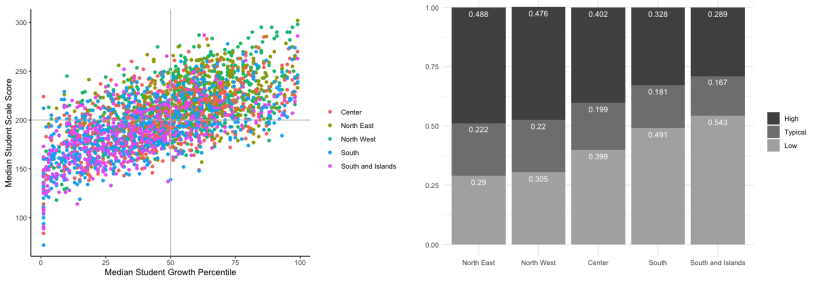

Fig. 6: School achievements (*left*) and growth (*right*) by geographical area.

6 Discussion and conclusions

Assessing students' skills is fundamental to improving the quality of the educational system. A way to determine the health of a school system is to constantly monitor and test the students' skills and knowledge to assess their level of preparation. This paper has adopted a longitudinal approach that gives important results and contributes significantly to the debate on assessing student performances. It allowed us to identify groups of particularly fragile students that should represent the target of ad hoc educational policies. Accounting for the non-vertically scaled nature of the INVALSI mathematics test scores at different grades, two approaches have been followed with different purposes. A three-level growth model has been estimated to account for the hierarchical structure typical of educational data. The impact of student-and school-specific covariates has been evaluated, but from this model, no proper conclusions on individual growth over time were derived. Student growth percentiles, not affected by the scaling of the scores, have been estimated. The progress of the cohort of students from grade 8 to grade 13 has been evaluated

in normative terms as the percentile in the conditional distribution of the score at grade 13 given the score history in previous grades.

Despite the non-vertically scaled nature, the INVALSI data represent an important source of information since the analysis is carried out annually, at the population level, and to any degree. Important characteristics are collected at the individual and at the school levels. These covariates make it possible to thoroughly investigate the complexity of the Italian educational system at both socioeconomic and territorial levels. The results have shown that, at the individual level, there are important differences between groups of students. These disparities (of gender, socioeconomic status, origin, school and geographical areas) are already present in the eighth grade and persist in the subsequent years. The educational system does not appear to be able to reduce these inequalities. On the basis of the derived results, it seems essential to adopt targeted actions at the school level more than at the individual one. Starting from the lower secondary school, deserving students with unfavourable socioeconomic conditions should be valued and motivated to select an upper secondary school that gives them the opportunity to best express their skills. Furthermore, for students with lower levels of achievement, it would be important to act with support activities in the school to avoid future failures.

References

1. Abrantes P. (2001). Mathematical competence for all: options, implications and obstacles. *Educational Studies in Mathematics*. 47(2), 125-143.
2. Agasisti T., Ieva F., and Paganoni A. M. (2017). Heterogeneity, school-effects and the North/South achievement gap in Italian secondary education: evidence from a three-level mixed model. *Statistical Methods and Applications*. 26(1), 157-180.
3. Agasisti T. and Vittadini G. (2012). Regional economic disparities as determinants of student achievement in Italy. *Research in Applied Economics*. 4(2), 33-54.
4. Azzolini D., Schnell, P. and Palmer J. (2012). Educational achievement gaps between immigrant and native students in two new immigration countries: Italy and Spain in comparison. *The Annals of the American Academy of Political and Social Science*. 643(1), 46-77.
5. Bates D.M., Maechler M., Bolker B. and Walker S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*. 67(1), 1-48.
6. Betebenner D.W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*. 28(4), 42-51.
7. Betebenner D.W and Linn R.L. (2010). *Growth in Student Achievement: Issues of Measurement, Longitudinal Data Analysis, and Accountability.*. ETS report. Princeton NJ: ETS.
8. Betebenner D.W. (2011). *A Primer on Student Growth Percentiles*. mimeo.
9. Betebenner D. W., Iwaarden A. V. and Domingue B. (2012). SGP: An R package for the calculation and visualisation of student growth percentiles and percentile growth trajectories. [Computer software manual]. Available from http://cran.r-project.org/web/packages/SGP/index.html (Rpackageversion0.9-0.0)
10. Bratti M., Checchi D. and Filippin A. (2007). Geographical differences in Italian students mathematical competencies: evidence from PISA 2003. *Giornale degli Economisti e Annali di Economia*. 66, 299-333.
11. Briggs D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*. 50(2), 204-226.

12. Briggs D.C. (2017). Learning theory and psychometrics: room for growth. *Assessment in Education: Principles, Policy and Practice*. 24(3), 351-358,
13. Briggs D. C. and Domingue B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics*. 38, 551-576.
14. Briggs D. C. and Peck F. A. (2015). Using learning progressions to design vertical scales that support coherent inferences about student growth. *Measurement: Interdisciplinary Research, and Perspectives*. 13, 75-99.
15. Castellano K. E. and Ho A. D. (2013). *A Practitioner's Guide to Growth Models*. Washington, DC: Council of Chief State School Officers.
16. Coleman J., Campbell E., Hobson C., Mcpartland J., Mood A., Weinfeldand F. and York R. (1966). *Equality of educational opportunity*. Technical Report. US: Washington DC.
17. Contini D. and Cugnata F. (2020). Does early tracking affect learning inequalities? Revisiting difference-in-differences modelling strategies with international assessments. *Large-scale Assessments in Education*. 8, 14.
18. Contini D., Di Tommaso M. L. and Mendolia S. (2017). The gender gap in mathematics achievement: Evidence from Italian data. *Economics of Education Review*. 58, 32-42.
19. Contini D. and Grand E. (2017). On estimating achievement dynamic models from repeated cross-sections. *Sociological Methods and Research*. 46(4). 683-714.
20. Curran P. J., McGinley J. S., Serrano D., and Burfeind C. (2012). A multivariate growth curve model for three-level data. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf and K. J. Sher (eds.). *APA Handbook of Research Methods in Psychology*. 3(17), 335-358. Washington, DC: American Psychological Association.
21. Deaton A. (1985). Panel data from time series of cross-sections. *Journal of Econometrics*. 30, 109-126.
22. Giambona F. and Porcu M. (2015). Student background determinants of reading achievement in Italy. A quantile regression analysis. *International Journal of Educational Development*. 44, 95-107.
23. Goldschmidt P., Choi K., Martinez F. and Novak J. (2010). Using growth models to monitor school performance: comparing the effect of the metric and the assessment. *School Effectiveness and School Improvement*. 21(3), 337-357
24. Grilli L., Pennoni F., Rampichini C. and Romeo I. (2016). Exploiting TIMSS and PIRLS combined data: multivariate multilevel modelling of student achievement. *The Annals of Applied Statistics*. 10(4), 2405-2426.
25. INVALSI. (2019). Rapporto Nazionale Prove INVALSI 2019. Available at the following website: `https://www.invalsiopen.it` [April 02, 2021]
26. Koenker R. (2005). *Quantile regression*. Cambridge: Cambridge University Press.
27. Matteucci M. and Mignani S. (2021). Investigating gender differences in mathematics by performance levels in the Italian school system. *Studies in Education Evaluation*. 70, 1-12.
28. Moffitt R. (1993). Identification and estimation of dynamic models with a time series of repeated cross-sections. *Journal of Econometrics*. 59, 99-123.
29. Niss M. (2003). Quantitative literacy and mathematical competencies. In B. L. Madison and L. A. Steen (Eds.), *Quantitative literacy: Why numeracy matters for schools and National Council on Education and the Disciplines*. 215-220.
30. Niss M. and Hojgaard, T. (2019). Mathematical competencies revisited. *Educational Studies in Mathematics*. 102(1), 9-28.
31. OECD (2019a). *Education at a Glance 2019: OECD Indicators*. Paris: OECD Publishing https://doi.org/10.1787/f8d7880d-en.
32. Panero M. (2019). Un'analisi longitudinale dei dati INVALSI di matematica di una stessa coorte alla scuola primaria. *Working paper INVALSI*, n. 41, Roma. Available at the following link: `https://www.invalsi.it/download2/wp/wp41_Panero.pdf`.
33. Raudenbush S. W., and Bryk A. S. (1988). Methodological advances in studying the effects of schools and classrooms on student learning. *Review of Research on Education*. 15(1), 423-473.
34. Rock D.A., Pollack J.M. and Weiss M. (2004). *Assessing cognitive achievement growth during the kindergarten and first-grade years*. ETS RR-04-22. Princeton, NJ: ETS
35. Schneeweis N. (2011). Educational institutions and the integration of migrants. *Journal of Population Economics*. 24 (4), 1281-1308.

36. Seltzer M., Choi K., and Thum Y. M. (2003). Examining relationships between where students start and how rapidly they progress: using new developments in growth modelling to gain insight into the distribution of achievement within schools. *Educational Evaluation and Policy Analysis*. 25(3), 263-286.
37. Seltzer M., Frank K. and Bryk A. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to the choice of metric. *Educational Evaluation and Policy Analysis*. 16, 41-49.
38. Singer J. D., and Willett J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. London: Oxford University Press.
39. Skovsmose O. (1994). Towards a critical mathematics education. *Educational Studies in Mathematics*. 27(1), 35-57.
40. Skovsmose O. (2016). What could critical mathematics education mean for different groups of students? *For the Learning of Mathematics*. 36(1), 2-7.
41. Student S.R. (2020). Vertical Scales, Deceleration, and Empirical Benchmarks for Growth. *Educational Researcher*, forthcoming.
42. UNESCO (2017). *Cracking the code: Girls' and Women's Education in Science, Technology, Engineering and Mathematics (STEM)*. https://unesdoc.unesco.org/images/0025/00253 4/25347 9E.pdf.
43. Verbeek M. and Vella F. (2005). Estimating dynamic models from repeated cross-sections, *Journal of Econometrics*. 127, 83-102.
44. Von Davier A., Carstensen C.H. and Von Davier M. (2006). *Linking Competencies in Educational Settings and Measuring Growth*. ETS RR-06-12. Princeton, NJ: ETS.
45. Yen W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich, and P. W. Holland (Eds.). *Linking and aligning scores and scales*. New York: Springer. 273-283.
46. Willett J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education*. 15(1), 345-422.