Quality & Quantity (2026) 60:10711–10738
https://doi.org/10.1007/s11135-026-02655-8

Revisiting the geography of grading standards: Evidence from Italian provinces

Emanuele Fedeli[1]

Received: 30 April 2025 / Accepted: 13 February 2026 / Published online: 19 March 2026

Abstract
The discourse on public education often suggests that some schools or geographical areas, across space or time, apply more lenient grading standards, as indicated by discrepancies between teachers' assessments and standardized test scores. This narrative appears in studies across various countries, including the UK, US, and Italy, where such disparities are linked to socio-economic background and student competencies. In Italy, grading standards exhibit notable geographical differences. While prior research confirms these patterns, it faces key limitations, particularly in addressing theoretical and methodological challenges due to data constraints and assessing the sensitivity of regional grading disparities to contextual factors. This study leverages INVALSI-SNV data, tracking three cohorts of students who began 5th grade in 2011/2012, 2012/2013, and 2013/2014. It follows them through 8th grade, identifies provincial grading standards in Italian and Mathematics, and accounts for methodological challenges. While regional grading disparities persist, they are less pronounced than earlier research suggests. These patterns change when controlling for a vector of socio-demographic characteristics of provincial cohorts as they progress from 5 to 8th grade. Although a geographic gradient in grading standards exists, it does not perfectly align with the traditional North–South divide; rather, it is primarily shaped by the ability composition of students within each province, especially in Mathematics.

Keywords Teachers' marks · Academic performance · Grading standards · Geographical divide

1 Introduction

In recent years, the debate centered on student assessment methods has been central to pedagogical discourse, particularly given the significant impact that these methods can have on students' skills and educational choices (Brookhart 2015; Sun and Cheng 2014). While dis-

✉ Emanuele Fedeli
emanuele.fedeli@unimi.it

[1] Department of Social and Political Sciences, University of Milan, Via Conservatorio, 7, 20122 Milan, Italy

cussions continue regarding the adoption of practices aimed at measuring students' competencies, numerical assessment remains pivotal in the evaluation of students in contemporary educational systems (Brookhart 1993; Carlana 2019; Modica, Dardanoni, and Pennsi 2009). Compared to the extensive literature treating grading standards as a variable of interest, measured as the difference between standardized tests and teachers' marks (Betts and Grogger 2003; Figlio and Lucas 2004; Lievore et al. 2024), relatively little attention has been devoted to assessing the reliability of these estimates, whether they are accurately measured, and how they should be interpreted. Considering this, the present study seeks to reinvigorate this body of research by placing a greater emphasis on the identification process and ruling out certain related issues, with the aim of offering clear implications for the debate.

The first issue is mainly theoretical and revolves around the connections among relevant factors. Grading standards are often explained by using standardized test scores as outcomes and teachers' assessments as predictors, based on the underlying assumption that teachers' evaluations influence test performance (Betts and Grogger 2003; Figlio and Lucas 2004; Lievore et al. 2024). However, this assumption (Cinelli et al. 2022; Heckman 1999; Pearl 2009) contradicts a principle of reality since the data-generating process suggests that it is the test score that determines the teacher's evaluation, not the other way around. Recognizing this distinction is crucial for accurately estimating grading standards (Pearl 2009). The second issue relates to both measurement error bias and endogeneity. Standardized test scores may be measured with error and can also be influenced by whether students are taught by the same or different teachers. To establish grading standards, a researcher might analyze results from an entry-level test given in the first year of secondary school, along with the grades assigned by teachers for that same test. In contrast, standardized tests are often administered during the school year or at the end of an educational level; in this manner, test performance is influenced by the Pygmalion effect (Rosenthal and Jacobson 1968). Italy is a good example of this because standardized tests are administered at the end of middle school, making it difficult to assess grading disparities during the school cycle. The third issue is that grading standards may reflect the composition of the student body (Lievore and Triventi 2022; Westphal et al. 2016). Indeed, teachers may adjust their grading based on the incoming students' competency levels. While not all teachers may be sensitive to this, evidence suggests that teachers are responsive to the socioeconomic and racial compositions of their students (Aucejo et al. 2022; Zanga and De Gioannis 2023).

Italy represents an intriguing case study regarding grading standards because the debate over geographical gradients in grading standards is highly prominent, supported by a body of literature that could benefit from a more accurate and thoughtful identification of grading standards (Braga et al. 2014; Brunello and Cappellari 2008; Triventi and Argentin 2015). Furthermore, as in many educational systems, Italian teachers are free to establish their evaluation criteria independently of common national- or school-level guidelines. While some general recommendations may exist, teachers retain significant discretion, allowing for potential arbitrary variations in their evaluation practices.

We estimated grading standards in the eighth grade of lower secondary school using teachers' grades and standardized test scores administered at the same grade level. First, we comprehensively tested whether grading gaps are present, to what extent they exist, and whether they differ from the national average. We achieved this through tests and corrections for measurement errors that had previously been overlooked. Finally, we examined the extent to which these grading differences respond to students' incoming skills, testing whether the possible geographic gradient reflects an asymmetric distribution of skills across provinces that is attributable to contextual and historical factors shaping students' development before entering lower secondary school. The results indicate that differences in grading standards across Italian provinces are less pronounced than previously suggested. Notably, the North–South divide in grading practices appears much less robust than commonly assumed. In fact, much of the perceived grading generosity in Southern provinces can be attributed to lower average levels of incoming student competencies, particularly in mathematics. The findings suggest that regional differences in grading standards are more closely linked to the composition of student cohorts rather than the arbitrary leniency of teachers. These results imply that policy interventions should prioritize strengthening students' foundational skills rather than enforcing standardized grading criteria.

2 Literature review

2.1 Relevance of grading standards in learning environments

Grading serves a significant role in education, impacting the lives of students across all academic levels and it affects their relationship with many actors, such as teachers, parents, peers, and employers (Batruch et al. 2023; Bowers 2011; Gentrup et al. 2020). Indeed, teachers' marks serve as tangible indicators for students and their families, offering insights into educational attainment that might otherwise not be clearly visible. These marks wield considerable influence, shaping students' motivation, self-assurance, and dedication to their studies (Fleischmann et al. 2021). Research, particularly focused on the U.S., underscores the significant impact of grading standards on students' subsequent learning and skill development (Betts and Grogger 2003; Figlio and Lucas 2004). Grading essentially serves as more than a mere numerical assessment; it shapes educational trajectories and societal perceptions (Lievore et al. 2024). Beyond the confines of academia, teachers' final marks hold significance, acting as a benchmark for university admissions and career prospects. Moreover, employers often view them as a measure of an individual's skills and suitability for a given role (OECD 2012). It is crucial to recognize that teachers' grading standards are multifaceted, influenced not only by students' academic proficiency but also by factors such as effort, motivation, behaviour, and classroom characteristics. This interplay can sometimes lead to strategic grading decisions (Aucejo et al. 2022; Iacus and Porro 2011), such as favoring certain categories of students over others (Tyner and Gershenson 2020), simply accounting for certain compositional aspects (Westphal et al. 2016), or being more lenient in some schools than others (Tyner and Gershenson 2020). The logic of lenient or strict grading standards is not limited to schools but can also be applied to geographical areas. Systematic differences may exist at the regional level, implying that students in certain areas are subject to more lenient grading standards than others.

2.2 The North–South divide in Italy

The issue of grading standards, particularly concerning the North–South divide in Italy, has garnered public attention. Research suggests that southern regions tend to have lower grading standards (Triventi and Argentin 2015), with teachers being more generous in assigning grades, especially to high-performing students. This geographical variation is also evident at the school level since considerable heterogeneity exists in grading standards not only across provinces but also within schools in the same region (Lievore, Fedeli, and Triventi 2023). An examination of the previous literature revealed numerous findings indicating geographical variances in educators' assessments within the educational continuum, spanning from the elementary to university levels. Although many of these investigations exhibit a local orientation or utilize controlled field experiments, they remain pivotal in informing the Italian educational landscape. Some docimological investigations posit that educators may appraise student examinations disparately (Bolletta 2001). Notably, this discrepancy in evaluative judgments has surfaced in multiple contributions. Localized research conducted in Trentino in 1999, focusing on state examinations at the end of upper secondary education, revealed that educators applied relatively lenient grading during oral assessments for students near the pass/fail threshold following the written examinations (Argentin and Tamanini 2001).

Others have scrutinized the relationship between final teachers' marks in upper secondary education and standardized test performance for admission into medical faculties at universities (Sestito and Tonello 2011). Admission relies on a uniform, national distribution of the examination, based on general knowledge. A very low association between teachers' marks at the end of secondary school and these tests has emerged, providing indirect evidence of divergent grading standards across educational establishments and geographical regions. Within a similar regional setting, Iacus and Porro (2011) investigated educators' grading practices in lower secondary schools, revealing disparities in grading severity among teachers based on classroom composition. In fact, the authors found that the severity of grading standards depends on the socio-demographic characteristics of the classroom, such as social background, gender, and migration background. Gay and Triventi (2011) analyzed the variances in grading standards pertaining to reading comprehension using data from PISA 2009. Their findings revealed a decrease in evaluation stringency from Northern and Central regions to Southern Italy and the Islands. In Southern regions, educators tend to award passing marks more generously compared to their counterparts in the North, and this pattern persists even after accounting for potential differences in students' individual characteristics and school types.

2.3 Identification of grading standards

2.3.1 Current scholarship and issues

The primary method for examining teachers' grading standards involves comparing teachers' marks against blinded, standardized tests that assess the same subjects (Betts and Grogger 2003; Figlio and Lucas 2004; Lievore and Triventi 2022; Malouff and Thorsteinsson 2016). In Italy, standardized tests are designed to objectively assess students' competencies, making them a valuable reference point and a reliable indicator of students' academic skills. Previous scholarship conceptualizes the discrepancy between standardized test results and teachers' marks as an indicator of teachers' strictness or leniency (Betts and Grogger 2003).

In previous literature, the study and estimation of grading standards typically follow a common approach. Students are nested within schools, classrooms, or geographical areas, and standardized test scores—used as proxies for latent achievement—are regressed on students' grades, which proxy teachers' evaluations and may incorporate socio-emotional factors and soft skills. This strand of the literature implicitly assumes that test scores represent an objective measure of achievement. However, under this assumption, test scores should conceptually explain grades rather than the reverse. Several important aspects are often overlooked in this framework. First, while teachers' grades are based on a synthesis of multiple past evaluations and are therefore relatively reliable, standardized tests are administered at a single point in time. As a result, students' performance may be affected by transitory factors unrelated to their true ability, for instance, a bad day, adverse weather conditions, or events occurring at home. This introduces measurement error in test scores, which need not be random. In fact, it may be systematically correlated with socioeconomic background, as students from disadvantaged families are more likely to be exposed to adverse events that negatively affect test-day performance. Second, test scores themselves may be influenced by teachers, implying that they do not purely reflect students' latent achievement. Ignoring this issue risks treating test scores as reliable measures of competence when this assumption may not hold. Third, previous scholarship often estimates grading standards without adequately controlling for students' sociodemographic characteristics or baseline competence. Together, these limitations may lead to biased results, whose statistical significance is frequently not formally tested. From a substantive perspective, the primary interest lies in assessing whether certain areas exhibit systematically different grading standards. Crucially, this requires benchmarking local grading practices against a meaningful reference point, such as the national average. In light of this overview, the following sections discuss each of these aspects in detail.

2.3.2 Setting the proper theoretical relationship

The first issue is primarily theoretical and concerns the relationships between relevant variables. In the literature, grading standards are typically described using standardized tests and teachers' marks, implicitly assuming that the teachers' mark explains the test score. However, this is quite plainly incorrect (Pearl 2009); the correct theoretical relationship is the opposite, meaning that the test score explains the teacher's mark. Inverting the positions of the two variables is beneficial in two respects. Indeed, prior standardized test scores can be used as an instrumental variable to mitigate measurement error in standardized tests (e.g., the so-called "bad test day"). At the same time, this strategy helps address teacher-related biases, such as teacher-loop bias and the Pygmalion/Golem effects, whereby students' outcomes may be influenced by current teacher exposure.

2.3.3 Measurement error and teacher loop bias

Ineed, the second methodological issue is known as an extension of the Pygmalion effect (Rosenthal and Jacobson 1968), as defined by educational psychologists. This effect describes how student–teacher interactions are internalized by the student and impact their academic growth perspective. For instance, if the teacher's interaction is positive and their expectations are high, the student internalizes them and creates a self-fulfilling prophecy by studying more and striving for academic success. Conversely, if the teacher's expectations are negative, the student may enter a negative spiral with academic repercussions (Gentrup et al. 2020). As depicted in a directed acyclic graph (Fig. 1), tests and teachers' marks

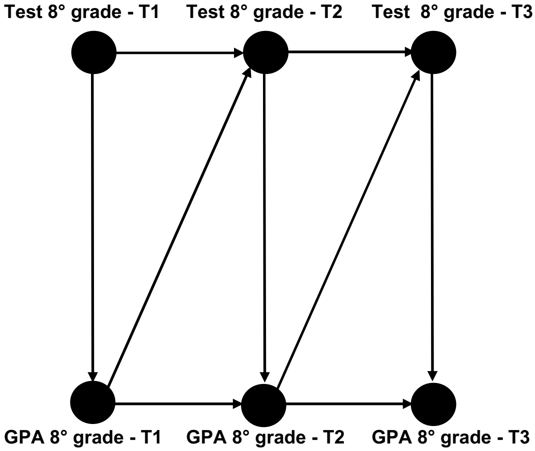

Fig. 1 Directed acyclic graph of teacher loop bias

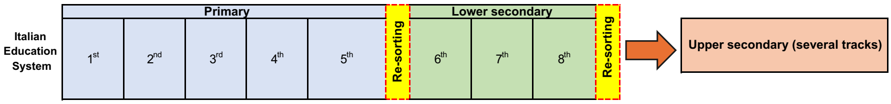

Fig. 2 Education system in Italy

influence each other over time in each grade, regardless of whether the researcher observes it or not. One possible solution is offered by a body of literature that proposes correcting current test scores with tests measured at different times to address the measurement error (Bradbury and Corak 2015; Jerrim and Vignoles 2013; Passaretta and Skopek 2021). This suggests that performance may simply be affected due to the test-taker's state on a particular day, or a school agency effect (i.e., endogeneity due to teacher–student interactions). This is relevant to our case because we leveraged the fifth-grade standardized test administered at the end of elementary school—before entry into middle school (Fig. 2)—to mitigate both measurement error and teacher bias, and to more accurately identify grading standards in eighth grade (Jerrim and Vignoles 2013). Teacher grades may be measured with some error; however, the grades reported in INVALSI are averages of written and oral assessments, where available, covering the period from September to January. As a result, the risk of measurement error is reduced. Moreover, these are not self-reported grades but administrative data drawn from official school registers.

2.3.4 Adjustment for socio-demographic and competence characteristics

The third methodological problem is a typical dilemma encountered when attempting to describe an association: which characteristics to control for. This study emphasizes the importance of controlling for basic socio-demographic characteristics and incoming students' competency levels during the transition from elementary to middle school. In this vein, it is reasonable to assume that teachers adjust their grading based on these levels. While not all teachers may be sensitive to this, evidence suggests that teachers are responsive to the socioeconomic and racial compositions of their students (Westphal et al. 2016).

As previously discussed, regarding "school agency," it is prudent to identify a proxy for academic competencies assessed prior to exposure to third-grade teachers, ideally during the elementary years. Therefore, we use standardized tests administered in fifth grade to measure the incoming cohort's ability level in each province, precisely a few months before entering middle school.

2.3.5 Testing the geographical divide

The fourth issue pertains to the variability in grading standards. To study a geographical gradient, as is the case here, it is necessary to assess the statistical discrepancy between the estimates and a reference mean. In this case, the reference mean represents the deviation of the geographical grading standard estimates from the Italian average. Otherwise, one would be unaware of the actual significant variation effectively observed. The four proposed solutions support the correct identification of grading standards and their accurate description, thereby enabling effective testing of a grading standard geographical gradient. Geographical comparisons that contrast different areas without accounting for baseline demographic characteristics and, most importantly, students' initial skill levels, are misleading. While this does not imply causal identification, the use of temporally ordered measurements between the fifth and eighth grade are good ways to address the geographical divide.

2.3.6 The Italian education system

Italy represents an ideal case study for grading standards because, both *de jure* and de facto, teachers have complete freedom in how they evaluate students. This leads to variability in grading across spatial units such as regions, as well as between schools and even within the same schools. In Italian primary and secondary schools, the Ministry of Education provides guidelines for grading students on a scale from 1 to 10, with 6 being the passing teacher's mark (Carlana 2019; Lievore et al. 2024). Students and families receive two report cards per academic year, typically in February and June, which detail students' academic progress. If a student receives a teacher's mark below 6 in any subject by the end of the year, they must retake an exam in that subject before the new school year begins in September.

The report cards typically display average teachers' marks for each subject based on various exams conducted throughout the semester. The structure and frequency of these exams vary depending on the subject, school regulations, and individual teacher preferences, giving teachers significant autonomy in their grading practices. While the Ministry of Education provides some guidelines, adherence to these guidelines varies among schools and teachers.

At the end of lower secondary schools, Italian students choose their high school track. Although teachers' marks and recommendations are not binding for track selection, students typically receive formal or informal advice and choose between vocational schools, technical schools, and lyceums. Lyceums, particularly traditional ones focusing on the humanities or science, are generally viewed as prestigious and often lead to university enrolment. In contrast, technical and vocational schools are geared towards entering the job market. Despite the formal stratification of upper secondary education in Italy, university enrolment is open to all high school graduates, regardless of their academic performance or final teachers' marks. However, some universities may require admission tests for certain programs (Contini and Triventi 2016).

3 Research design

3.1 Model specification for grading standards

The research design of this study is centered on estimating grading standards across Italian provinces and testing them against the Italian grading standard as a reference. The choice of provinces was made because it represents the smallest legal geographic entity, thereby allowing for a more detailed exploration of geographic differences. By adapting the method of Betts and Grogger (2003), teacher grading standards were estimated for each province; therefore, all students in the same provincial unit were assessed against the same grading standards. Relying on two separate regressions for mathematics and language, the grading standard estimates were obtained by regressing students' GPAs (teachers' mark point average) in the respective subjects on test scores in mathematics and language. Additionally, we added a vector of provincial fixed effects, which represents our estimates of grading standards and we did not adopt cohort fixed effects to ensure a sufficient and robust sample size for each province.

We adopted three specifications to address the points raised in the literature review section. In the primary specification, we reversed the estimation of grading standards when compared to previous literature and estimated equation A (hereafter, Model A). Subsequently, to account for the teacher loop bias, we estimated equation B (hereafter, Model B), instrumenting the eighth-grade teachers' mark test score with that of the fifth-grade teachers' mark. Finally, in the last specification (C), we introduced individual-level controls (e.g., sex, ethnic background, and socioeconomic background) and provincial-level controls, including the same socio-demographic factors, as well as the provincial cohort's prior achievement measured in the fifth grade (hereafter, Model C).

(1) (Model A) $GPA_{i,p,t_0} = \alpha_0 + \theta_0\ Prov\ FE_p + \beta_0 Test_{i,p,t_0} + \varepsilon_{i,p,t_0}$

(2) (Model B) $GPA_{i,p,t_0} = \alpha_0 + \theta_0\ Prov\ FE_p + \beta_0 Test_{i,p,t_0} + \varepsilon_{i,p,t_0}$

$$\text{IV: } Test_{i,p,t_0} = \alpha_0 + \theta_0\ prov\ FE_p + \beta_0 Test_{i,p,t_{-3}} + \varepsilon_{i,p,t_0}$$

(3) (Model C) $GPA_{i,p,t_0} = \alpha_0 + \theta_0\ Prov\ FE_p + \beta_0\ Test_{i,p,t_0} + \beta_1 X_{i,p,t_{-3}} + \beta_2 Z_{p,t_{-3}} + \varepsilon_{i,p,t_0}$

$$\begin{aligned}\text{IV: } Test_{i,p,t_0} = \alpha_0 + \theta_0\ prov\ FE_p + \beta_0 Test_{i,p,t_{-3}} \\ + \beta_1 X_{i,p,t_{-3}} \beta_2 Z_{p,t_{-3}} + \varepsilon_{i,p,t_0}\end{aligned}$$

where GPA_{i,p,t_0} is the average mathematics or language score in the eighth grade; $\theta_0 Prov\ FE_p$ is the fixed effect at the provincial level for measuring the grading standard; $\beta_0\ Test_{i,p,t_0}$ is the standardized test score in the eighth grade; $\beta_1 X_{i,p,t_{-3}}$ is a vector of individual socio-demographic characteristics; $\beta_2 Z_{p,t_{-3}}$ is a vector of the provincial socio-demographic characteristics of the incoming cohorts collected in the fifth grade, which includes provincial prior achievement; $\beta_0\, Test_{i,p,t_{-3}}$ is the standardized test score from the fifth grade used as instrumental variable to control for teacher loop bias; and ε_{i,p,t_0} is the error term. We anticipate that the inclusion of socio-demographic controls at the individual or provincial level, whether jointly or separately, does not alter the conclusions of this study—namely, the pivotal role of the cohort's incoming competence at the provincial level.

As described in the model specification section in Model A, we constructed grading standards using only teachers' GPAs and standardized test scores measured in the eighth grade. In Model B, we introduced an instrumental variable (IV) approach by adjusting the eighth-grade test scores at the individual level using the results of previous standardized test score taken in the fifth grade. Conceptually, this instrument is valid because it serves as a prior measure of the student's achievement, thereby helping to correct potential measurement error and endogeneity arising from the teacher–student relationship. The correlation between the two test scores was high (0.52 in Language and 0.43 in Mathematics), and the model exhibited a strong first-stage F-statistic (above 300 in both subjects), indicating instrument relevance. Hence, in the first stage, current test scores were regressed on past test scores and a set of individual- and family-level covariates. The predicted values from this regression were then used in the second stage to estimate the effects on the outcomes of interest. A critical point arose in Model C, where we introduced individual provincial-level controls, namely the provincial cohort's prior achievement, measured by the average fifth-grade standardized test scores.

One potential point of concern is that the fifth-grade test is used both as an individual-level instrument in Model B and later reintroduced as an aggregate provincial-level control in Model C. However, several considerations mitigate this concern. First, while alternative indicators of prior achievement (such as second-grade standardized tests) could be considered, the fifth-grade assessment is the most appropriate measure, as it captures student achievement at the end of primary education, immediately before the transition to lower secondary school. Second, the correlation between individual fifth-grade test scores and their provincial aggregates is low (0.14 in language and 0.11 in mathematics; see appendix, Figs. 7 and 8), substantially reducing concerns about multicollinearity. Third, the main results remain unchanged even when the IV correction for eighth-grade test scores is omitted in Model C. Moreover, measurement error arising from transitory factors (e.g., particularly good or bad test days) is not systematically correlated across students within provinces, and the provincial average is therefore independent of the individual-level measurement error. Under these conditions, the instrument remains valid, and the 2SLS estimates are consistent and unbiased, allowing us to effectively correct for measurement error.

We ran ordinary least squares (OLS) for each previously discussed equation (A, B, C). We then estimated the equations with our analytical sample at the individual level and stored the estimates for each province (see appendix Tables 3, 4 and 5). When estimating equations B and C, we adopted an instrumental variable approach (Jerrim and Vignoles 2013). The coefficients of province fixed effects represent the estimated grading standards in language and mathematics at the provincial level. This implies that if variation across grading exists, a province with a lower θ has stricter grading standards. If $\theta_1 < \theta_2$, a student in province 1 is exposed to stricter grading standards relative to a student in province 2; the two students have an equal GPA in subject s, but the student in province 1 has a higher test score in subject s than the student in province 2.

We reversed the polarity of the variables so that a higher score along the grading standard distribution was associated with greater severity (simply reversing $\theta_1 < \theta_2$, in the previous example). The second step was to normalize the scale between 0 and 100 solely for map visualization, where 100 represents maximum severity, and 0 represents extreme leniency. Additionally, as an engagement rule for the graphical visualization of grading standards among provinces, we constructed five grading standard bands: 0, 0.20, 0.40, 0.60, 0.80, and 1. This is a reasonable choice that minimizes the strategic use of multiple bands that could exacerbate marginal provincial differences without substantive justification and create a visual fallacy.

The statistical relevance should be examined from two perspectives: model-based and geographical. The model-based perspective evaluates how many spatial units change their grading severity depending on the specified model. The geographical perspective assesses whether the estimate for each spatial unit (i.e., a region in our case) significantly differs from the national grading average within each model. To test these relevancies, we employed the Paternoster test and evaluated the significance of spatial units across different model specifications. Specifically, we compared the differences between Model A and Model B, and between Model B and Model C, using the Paternoster test. The results were visualized using a modified Sankey diagram to illustrate changes in grading severity. Regarding the geographical perspective, we assessed the significance of spatial estimates compared to the Italian national average. To calculate the national average from the provincial estimates, weighted means were used:

$$\widehat{GS}_{average} = \frac{\sum_{i=1}^{n} w_p * GS_p}{\sum_{i=1}^{n} w_p}; \widehat{SE}_{average} = \frac{\sum_{i=1}^{n} w_p * SE_p}{\sum_{i=1}^{n} w_p}; w_p = \frac{1}{SE_p^2} \tag{1}$$

where GS_p is the grading standard estimate in province p and $SE_p$is the standard error in the province p. The weighted mean of the estimates was calculated by assigning greater importance to estimates with smaller standard errors since these are considered more precise (Hedges and Olkin 2014).

3.2 Data and variables

We use data from the INVALSI- SNV dataset, which provides comprehensive information on student competencies and educational trajectories across various grading levels, covering the entire student population. We followed three consecutive cohorts (2011/2012, 2012/2013, and 2013/2014) to construct a panel dataset of approximately 800,000 students enrolled from the fifth to eighth grades. Specifically, we used standardized tests and teacher-assigned grades in the eighth grade to estimate grading standards in mathematics and language. To address the previously discussed issues, we leveraged fifth-grade standardized test scores to mitigate the teacher loop bias. Additionally, we incorporated socio-demographic variables such as sex, ethnic background, and socioeconomic status to explore the factors driving geographic gaps.

The broad coverage of the INVALSI-SNV dataset enables a thorough examination of territorial discrepancies, encompassing a diverse range of individuals across provinces. Par-

Table 1 Descriptive statistics of the individual dataset

	MEAN	SD	MIN	MAX
Individual				
Female	0.49			
Migration background	0.12			
ESCS	0.00	1.00	− 3.28	2.72
Test (eighth grade)—Language	0.00	1.00	− 5.88	4.51
Test (eighth grade)—Mathematics	0.00	1.00	− 4.84	4.37
Teacher mark—Language	0.00	1.00	− 5.18	2.49
Teacher mark—Mathematics	0.00	1.00	− 4.28	2.24
Test (fifth grade)—Language	0.00	1.00	− 5.72	4.43
Test (fifth grade)—Mathematics	0.00	1.00	− 5.33	4.30
Provincial				
Share of girls	0.50			
Share of non-natives	0.12			
ESCS	0.00	1.00	− 1.94	2.25
Test (fifth grade)—Language	0.00	1.00	− 2.62	1.51
Test (fifth grade)—Mathematics	0.00	1.00	− 2.53	2.17

All continuous measures are standardized

Table 2 Descriptive statistics of the provincial grading standard estimates

Grading standard	MODEL	MEAN	SD	MIN	MAX
Language	A	0.002	0.111	− 0.263	0.267
	B	− 0.010	0.144	− 0.321	0.370
	C	0.041	0.102	− 0.216	0.307
Mathematics	A	− 0.005	0.091	− 0.193	0.187
	B	− 0.019	0.138	− 0.326	0.239
	C	0.032	0.074	− 0.208	0.262

ticipant numbers vary widely, from 1,100 students in Aosta province to 92,869 in Milan province. We excluded the province of Bolzano due to limited observations, reflecting its lower participation in the surveys during the study period. To ensure sufficient and robust sample sizes for each province, we pooled data across all three cohorts without distinguishing between them. The main variables were drawn from school administrative records, standardized tests conducted at specific grade levels, and questionnaires administered to all students. Our initial dataset (Table 1) is at the individual level, and we employed both individual- and provincial-level characteristics to estimate grading standards (Model A, Model B, and Model C). After obtaining the estimates, we switched to a provincial-level dataset (Table 2). We do not report values for each province individually since our goal is not to rank provinces but to explore broader geographical patterns. The outcome variable is the GPA collected in eighth grade for both mathematics and language, based on oral and written examinations. We computed a simple arithmetic average of the two subjects. The primary independent variables are the standardized test scores in mathematics and language at the eighth grade, where higher scores reflect better student performance. In terms of model specification, we controlled for key provincial-level socio-demographic factors such as gender composition and migration background (distinguishing between natives and first- or second-generation migrants).

4 Results

4.1 Descriptives and main results

We discuss the main results from the three estimations of grading standards in each Italian province (Fig. 3), as discussed in the previous section on the research design (see appendix - tables 3, 4, and 5 for the overview of all estimates). The first estimation (Model A) follows the Italian and international literature, exploiting the discrepancy between teachers' marks and test scores, but with a simple contribution by reversing the equation to better align with the data-generation process, namely that teachers' marks depend on the test and not vice versa. As evidenced by the maps provided for language results, Italy is characterized by relatively high grading standards in several provinces in the Northeast and parts of the Center. However, a clear North–South divide is not evident since many provinces in the North, Center, and South fall within the same band, except for a subset of provinces in Southern Italy, which cluster in the lowest band.

Although the results only partially confirm previous contributions in the field (Triventi and Argentin 2015), it is important to note that our research design differs in terms of multiple cohorts of students within provinces included.

The estimations started to change when we corrected for the teacher loop bias (Model B) present in the INVALSI-SNV dataset. As previously discussed, the concomitant stan-

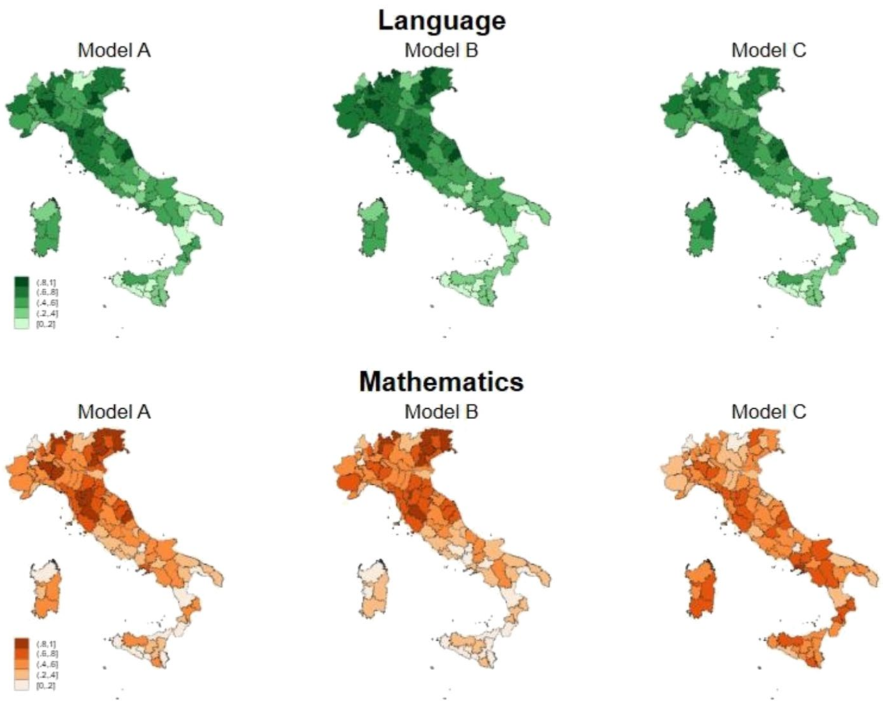

Fig. 3 Grading standards at the provincial level by model specifications in language and mathematics. *Note*: Model A: regressing GPA on standardized test; Model B: prior standardized test results collected in fifth grade as the instrumental variable; Model C: adjusted for socio-demographic characteristics and incoming competence level (proxied by fifth grade standardized test results)

dardized eighth-grade test may be biased by exposure to the same teacher for 3 years and the consequent Pygmalion effect (Rosenthal and Jacobson 1968). Some areas in the Central-North and South have different estimations, while the remainder of Italy remains largely constant. Even more relevant is what occurred when we investigated the third map (Model C), where we not only used the correction for teacher loop bias, but also included a set of basic socio-demographic variables (at the individual and provincial levels) and students' average ability upon entry into middle school. There is a clear shift between the bands, especially for the provinces in Southern Italy. We stress the fact that this change is essentially due to the inclusion of the average ability, which is a message of great relevance. Indeed, when controlling for the competencies of incoming cohorts, grading standards in the South become "less" lenient (comparison: Model C vs Model B), suggesting that teachers are more generous simply because they deal with less competent students, leading them to provide more accommodating evaluations (Westphal et al. 2016). Such results are also confirmed in mathematics, for which they are even more pronounced. Indeed, the higher bands are mainly present in Central Italy, while other Italian provinces fall within the lower bands (Model A). When we begin to correct for the teacher loop bias and measurement error (Model B), some territorial changes are more pronounced in Southern Italy, especially along the Tyrrhenian coast. However, when we introduced the set of control variables, the situation dramatically reversed. Southern Italy was assigned to medium and high grading standard bands, implying that, even more than for language, the average ability level explains the generous grading practices in this region. The message is quite clear when observing both main subjects of the education system. The notion that teachers in Southern Italy are more generous, as often portrayed, was not immediately supported when we started correcting the test for measurement error and teacher loop bias. The narrative of teachers being more generous in the South is discredited when we compare provinces while keeping basic socio-demographic factors and average ability level constant. This indicates that the real point of discussion in public debate should not be whether teachers in the South have more generous proclivities when evaluating students, but rather that the stock of students is lower in terms of their ability when enrolling in middle school. From this perspective, teachers merely align their grading to the average level of competence that students have when entering schools. If teachers were genuinely overgenerous, regardless of context, it would call for interventions such as guidelines or awareness campaigns on how to evaluate students. This may be not the case since teachers simply assess students based on their competence levels upon entrance into middle school.

4.2 Testing model differences within provinces

We tested whether significant differences exist between Models A, B, and C for each subject (Fig. 4). Both tests are important because they help avoid drawing conclusions based on artificial differences. For the first test, significant differences were calculated using a Paternoster test, and a bar chart was chosen to provide a clear graphical representation. The first two bars examine whether the transition from Model A to Model B results in a significant difference in the estimates, effectively testing whether any meaningful change exists or if the estimates remain virtually identical. As shown, in both Italian and mathematics, the green bar reaches 50%. This implies that, of the provincial estimates in Model A, at least half show significantly different estimates in Model B. These findings indicate that correction with IVs has an effect. Using the same reasoning, we analyzed the other two bars, particularly the purple bar, which approaches 70% in both subjects. This suggests that the switch from Model B to Model

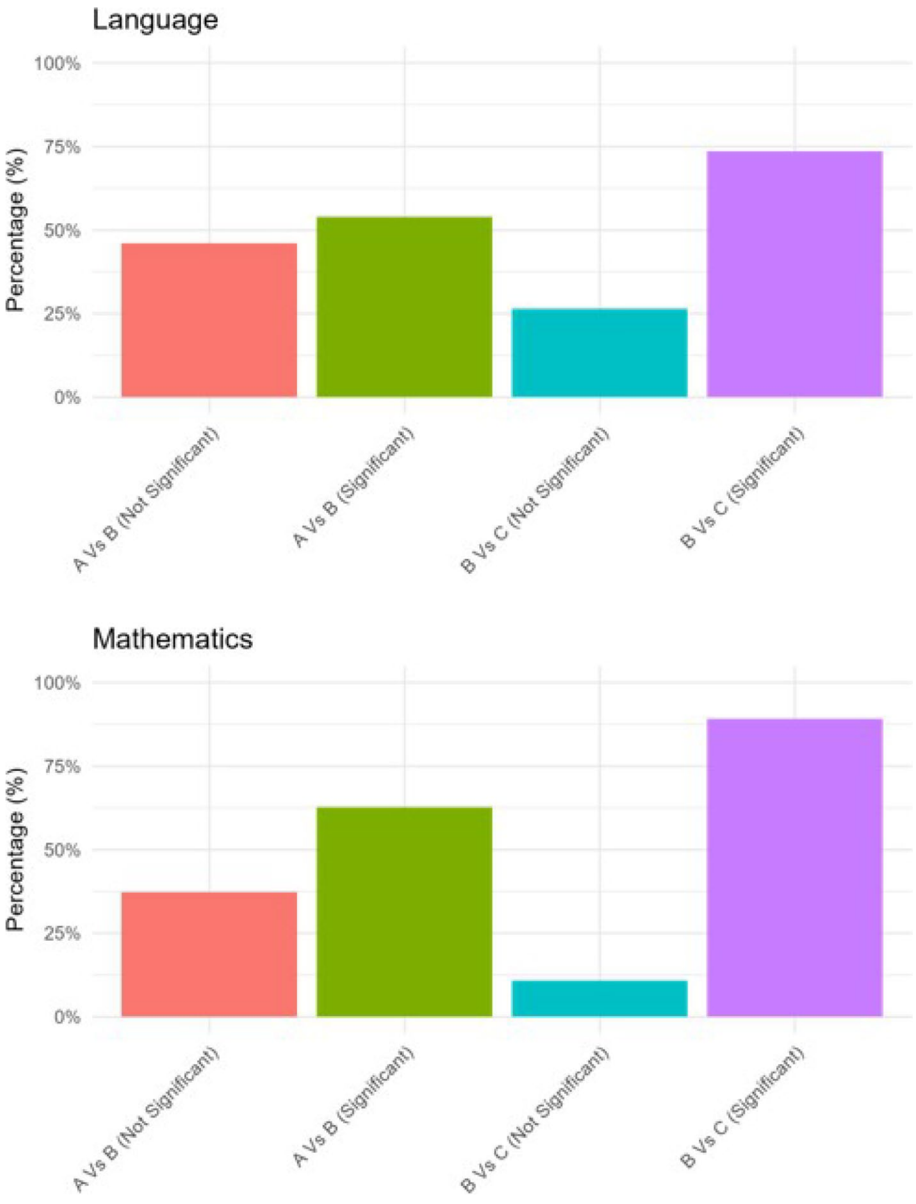

Fig. 4 Changes in significance between models across subjects. *Note*: The plot accounts for the significance of difference estimates across models, namely A versus B and B versus C. Orange indicates the proportion of estimates that do not exhibit significant differences between Models A and B. Green indicates the proportion of estimates that exhibit significant differences between Models A and B. Blue indicates the proportion of estimates that do not exhibit significant differences between Models B and C. Purple indicates the proportion of estimates that exhibit significant differences between Models B and C

C is decisive since controlling for basic socio-demographic variables and individual-level competencies leads to significant differences. This provides an alternative perspective to the map-based analysis discussed earlier, but with the added advantage of statistically testing these differences.

4.3 Testing geographical differences

To analyze geographical differences using Italy as a reference (Figs. 5 and 6), we plotted the average national value of grading standards against each provincial estimate for every model specification. For Model A—where the equation is reversed, as previously discussed—we used the color red. For Model B, which incorporates an IV, we used green. For Model C, we used purple. The significance of each difference between a provincial estimate and the national average was tested using the Paternoster test. Filled markers denote statistically significant differences, while empty markers indicate non-significant differences. On the x-axis, the abbreviations of Italian provinces are displayed. These are not arranged alphabetically but are ordered according to the magnitude of the estimates. The province labels are color-coded to reflect their geographical location (i.e., North, Center, or South).

Upon investigating the panels for both subjects (Italian and Mathematics), it becomes clear that the different model specifications substantially affect the estimates since the positions of

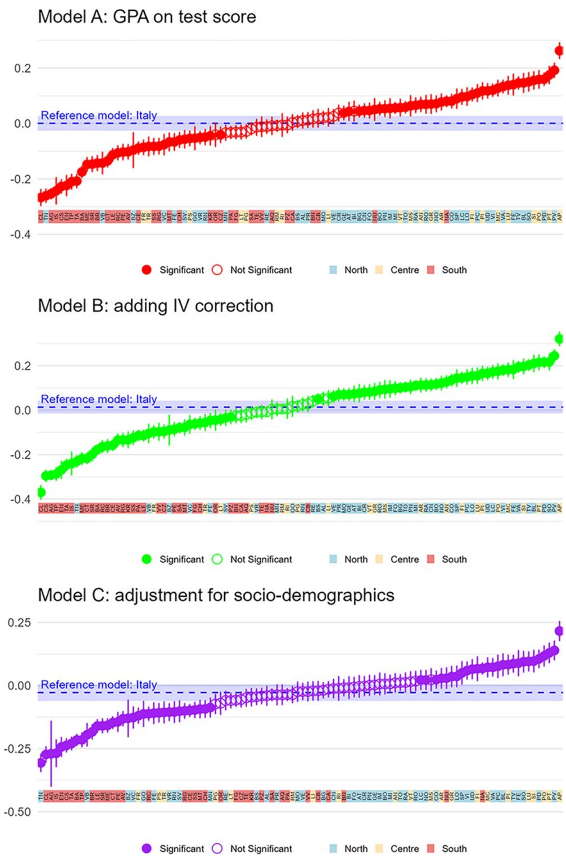

Fig. 5 Testing provincial and national differences in Language across the model. Note: Model A: regressing GPA on standardized test score; Model B: previous standardized test score collected in fifth grade as the instrumental variable; Model C: adjusting for provincial socio-demographic characteristics and incoming competence level (proxied by fifth grade standardized test). The reference model is Italy, computed as described in Sect. 3

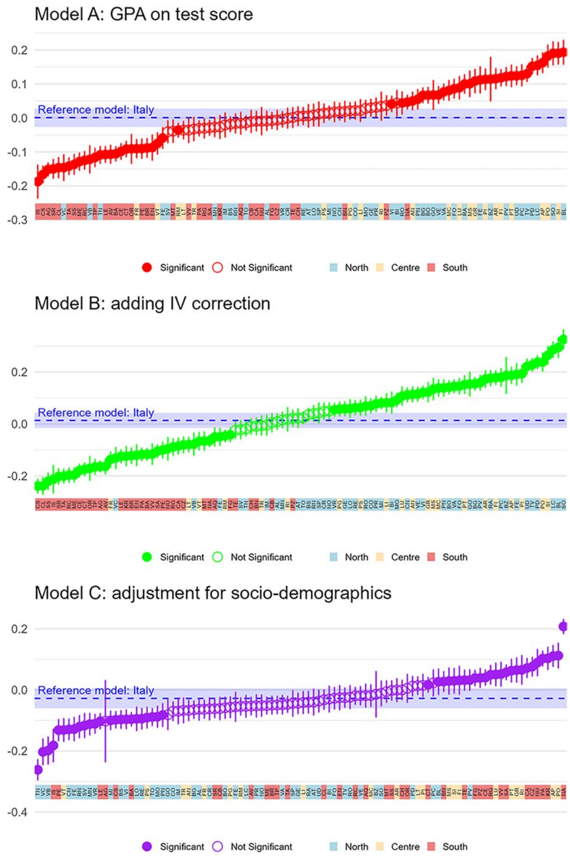

Fig. 6 Testing provincial and national differences in Mathematics across models. Note: Model A: regressing GPA on standardized test score; Model B: adopting previous standardized test score collected in fifth grade as the instrumental variable; Model C: adjusting for provincial socio-demographic characteristics and incoming competence level (proxied by fifth grade standardized test). The reference model is Italy, computed as described in Sect. 3

the provinces shift and, more importantly, the significance of certain provinces changes. This is particularly evident in Model C, where we controlled for basic sociodemographic characteristics (sex, migration background, and socioeconomic origin) at the individual level and initial competency at the provincial level. The visualization highlights that many provinces—in both mathematics and Italian—do not significantly differ from the national average. As shown along the x-axis, many southern provinces (indicated in red) begin to shift in Model B, with more substantial shifts observed in Model C, particularly for mathematics.

This pattern reinforces the broader takeaway of this study: When controlling for provincial differences in student competency levels, the geographical gradient in grading standards diminishes considerably. Although this reduction is evident in both subjects, it is more pronounced in mathematics.

A delicate aspect of this work concerns the vector of controls at both the individual and provincial levels, and how it interacts with the use of the IV. As previously explained, this setup is valid because the fifth-grade test score was used as an IV, while the same test score, averaged at the provincial level, served as a control variable. Alternative specifications are possible; for example, by omitting the IV and directly controlling for the average fifth-grade test score at the provincial level. In Figs.9 and 10 (see Appendix), we present the results of a robustness check in which we solely controlled for the average incoming competence of cohorts, without including any individual-level or additional provincial-level controls. The pattern for mathematics remained nearly identical, with little change observed in the spatial distribution of provinces; for example, the provinces in the South (highlighted in red) continued to stand out. In contrast, for Italian, the situation is somewhat different; however, this aligns with earlier findings suggesting that the grading standards for Italian are less sensitive to socio-demographic controls when compared to mathematics. What clearly emerges is a degree of heterogeneity between the two subjects. It is likely that including both individual and provincial level characteristics, especially average ability and social background at the provincial level, better captures the socio-demographic composition of cohorts. In this study, the key takeaway is that contextual factors must be considered: When controlling for provincial characteristics, particularly incoming competence, the North–South divide becomes less robust and less clear, especially for mathematics. Another important methodological point concerns the comparison between the inverted equation used in our study and the non-inverted specification commonly employed in previous literature. Such comparisons are not meaningful for several reasons. We adopt the inverted specification because it is the more appropriate choice from a theoretical perspective, an approach originally proposed in earlier studies but later overlooked. Moreover, the inversion allowed for the use of an IV to correct for measurement error. Ultimately, comparing results from inverted and non-inverted equations offers limited insights since the underlying variables are the same and highly correlated.

5 Conclusions

Educators conscientiously consider students' initial competencies, thereby supporting the notion that grading standards are sensitive to learning environment composition (Lievore and Triventi 2022; Westphal et al. 2016). From a regulatory perspective, student evaluations should certainly be "standardized" across the national territory on a scale from 0 to 10, with centrally approved scores corresponding to a level of subject knowledge; in this case, mathematics and language. However, it is equally true that the Italian school system acknowledges teachers' absolute independence in evaluating students and the free decision to weigh evaluations in light of students' progress (Bowers 2011; Chetty et al. 2014a, 2014b). Therefore, it is reasonable for this teacher–student relationship to be evident at a geographic level if the corresponding context exhibits low incoming competency. These guidelines for managing grading standard models can also be applied at the school level, with significant policy implications. Notably, such an approach would allow for the identification of which schools differ in their grading practices and an exploration of the potential reasons behind these differences.

As previously stated, this work does not aim to establish causal identification but instead suggests that the quantification of grading standards requires careful reconsideration.

Furthermore, this pattern is not dissimilar from another important finding observed in the Italian labor market, where, in the face of company layoffs, courts are much more inclined to reinstate workers in labor markets with high unemployment rates (Ichino et al. 2003). This is because judges are aware that confirming dismissal in a market with high unemployment would greatly diminish one's likelihood of being hired in the future. The same reasoning can be applied in our case. Not only do teachers assess the level of incoming competency, but they may also be slightly more lenient due to the local context. Future research can follow this path by improving both the estimation of grading standards and seeking to understand the drivers of local contexts or the cohort characteristics leading to these geographical disparities.

Appendix

See Appendix Figs. 7, 8, 9, 10 and Tables 3, 4 and 5.

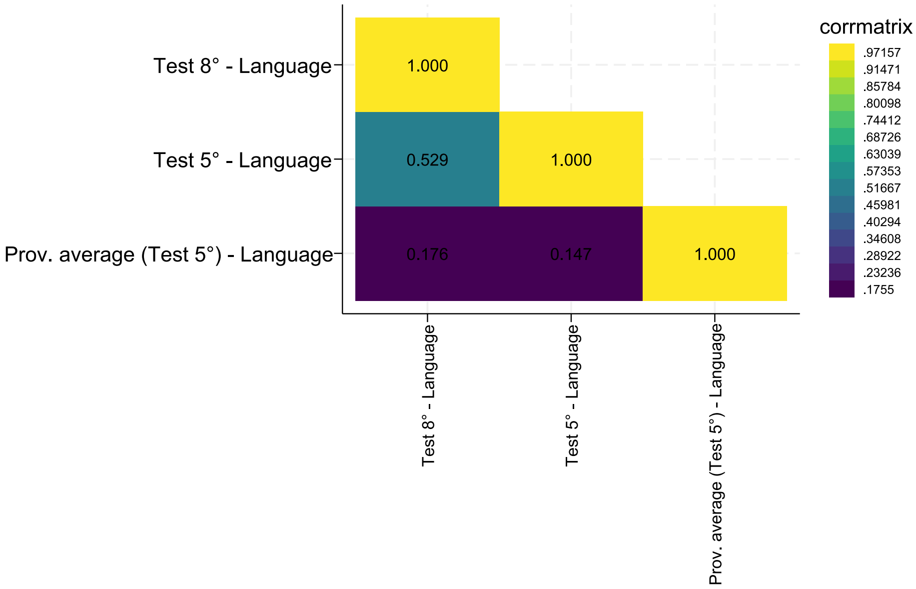

Fig. 7 Matrix correlations for language

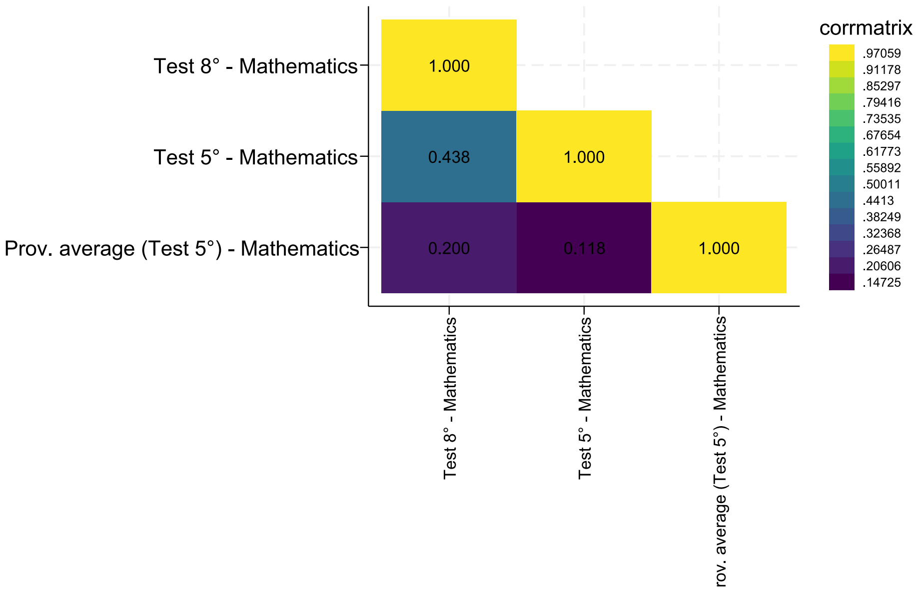

Fig. 8 Matrix correlations for mathematics

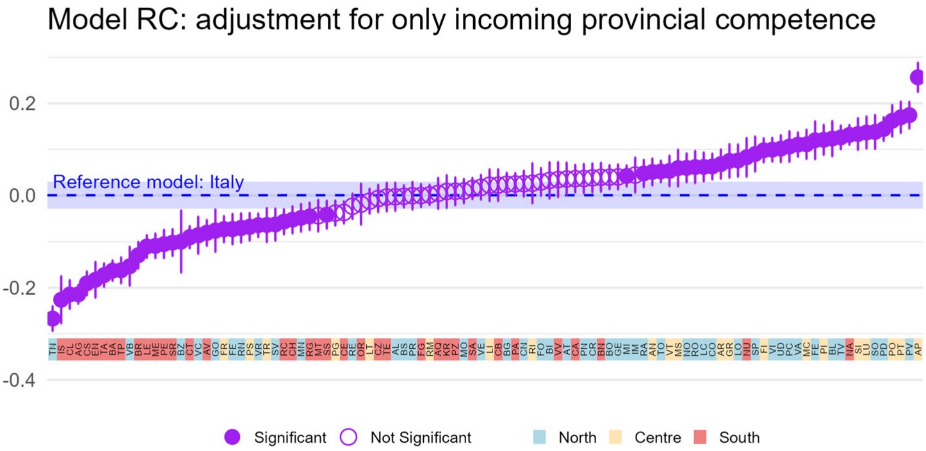

Fig. 9 Testing province and national difference across model in Language only adjusting only for incoming competence at provincial level. *Note*: Model RC: adjusting for incoming competence level (proxied by 5th grade standardized test). The refence model is Italy computed as described in Sect. 3

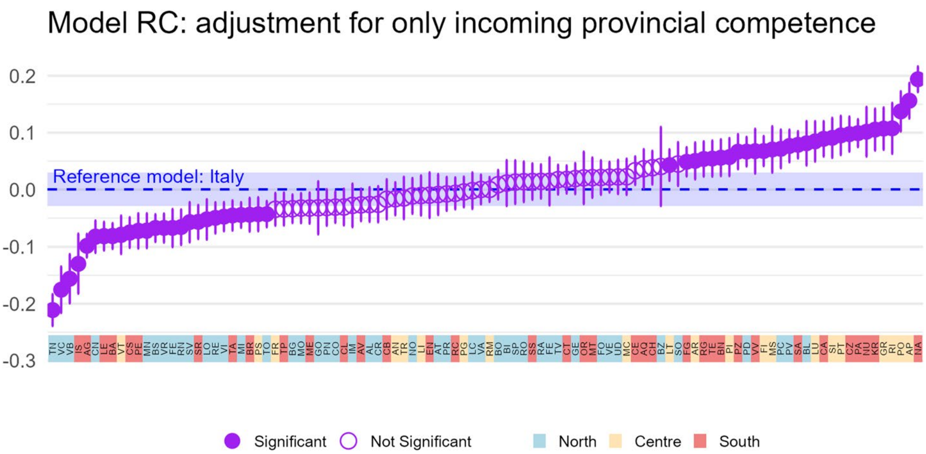

Fig. 10 Testing province and national difference across model in Mathematics only adjusting only for incoming competence at provincial level. *Note*: Model RC: adjusting for incoming competence level (proxied by 5th grade standardized test). The refence model is Italy computed as described in Sect. 3

Table 3 Estimates of provincial grading standards by model across subjects 1/3

	Language						Mathematics					
	Model A		Model B		Model C		Model A		Model B		Model C	
Prov	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*
AG	0,253	0,009	0,292	0,010	0,270	0,065	0,152	0,010	0,164	0,010	0,102	0,068
AL	− 0,011	0,014	− 0,052	0,015	0,040	0,019	0,009	0,014	− 0,009	0,015	0,064	0,018
AN	− 0,069	0,013	− 0,127	0,014	− 0,011	0,019	− 0,050	0,013	− 0,114	0,014	0,067	0,017
AP	− 0,263	0,014	− 0,321	0,015	− 0,216	0,019	− 0,162	0,015	− 0,188	0,016	− 0,111	0,017
AQ	0,008	0,016	0,020	0,017	0,037	0,020	0,016	0,016	0,050	0,018	0,019	0,019
AR	− 0,078	0,015	− 0,113	0,016	− 0,028	0,020	− 0,115	0,015	− 0,173	0,016	0,001	0,019
AT	− 0,044	0,016	− 0,074	0,018	0,004	0,022	− 0,006	0,017	− 0,022	0,018	0,040	0,022
AV	0,102	0,013	0,135	0,014	0,131	0,014	0,107	0,013	0,164	0,015	0,053	0,015
BA	0,175	0,010	0,177	0,011	0,216	0,011	0,106	0,011	0,117	0,011	0,096	0,011
BG	− 0,045	0,011	− 0,101	0,012	0,019	0,015	− 0,066	0,011	− 0,155	0,012	0,065	0,014
BI	− 0,045	0,018	− 0,076	0,020	0,008	0,022	− 0,043	0,019	− 0,082	0,020	0,030	0,022
BL	− 0,146	0,017	− 0,211	0,019	− 0,082	0,022	− 0,193	0,018	− 0,295	0,019	− 0,027	0,021
BN	− 0,016	0,015	0,025	0,016	0,009	0,016	− 0,019	0,015	0,009	0,016	− 0,027	0,017
BO	− 0,053	0,011	− 0,118	0,012	− 0,007	0,020	− 0,067	0,012	− 0,141	0,013	0,058	0,018
BR	0,143	0,014	0,161	0,015	0,183	0,014	0,088	0,014	0,122	0,015	0,050	0,015
BS	− 0,007	0,011	− 0,051	0,011	0,044	0,016	0,019	0,011	− 0,036	0,012	0,097	0,015
CA	− 0,004	0,014	0,025	0,015	0,015	0,015	0,009	0,015	0,085	0,016	− 0,072	0,016
CB	− 0,019	0,016	− 0,028	0,017	0,021	0,019	0,010	0,016	− 0,005	0,018	0,061	0,019
CE	0,089	0,011	0,157	0,012	0,099	0,012	0,099	0,012	0,180	0,012	− 0,040	0,013
CH	0,059	0,014	0,058	0,015	0,093	0,017	− 0,005	0,014	0,009	0,015	0,000	0,016
CL	0,267	0,015	0,370	0,016	0,273	0,015	0,146	0,015	0,238	0,016	0,032	0,016
CN	− 0,047	0,012	− 0,114	0,013	0,012	0,018	− 0,018	0,013	− 0,112	0,014	0,129	0,017
CO	− 0,081	0,012	− 0,136	0,013	− 0,023	0,017	− 0,022	0,013	− 0,070	0,014	0,070	0,015
CR	− 0,038	0,014	− 0,080	0,015	0,002	0,019	0,001	0,015	− 0,049	0,016	0,062	0,018
CS	0,227	0,012	0,296	0,013	0,237	0,013	0,166	0,012	0,239	0,013	0,098	0,014
CT	0,135	0,011	0,217	0,012	0,147	0,012	0,092	0,011	0,174	0,012	− 0,017	0,012

Table 3 (continued)

	Language						Mathematics					
	Model A		Model B		Model C		Model A		Model B		Model C	
Prov	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*
CZ	0,041	0,014	0,096	0,015	0,055	0,015	0,006	0,014	0,081	0,016	− 0,076	0,016
EN	0,224	0,018	0,269	0,020	0,243	0,019	0,087	0,019	0,120	0,020	0,025	0,020
FE	0,067	0,015	0,046	0,016	0,109	0,021	0,059	0,016	0,049	0,017	0,128	0,019
FG	0,031	0,012	0,061	0,013	0,058	0,012	0,008	0,012	0,042	0,013	− 0,039	0,013
FI	− 0,105	0,012	− 0,146	0,013	− 0,067	0,024	− 0,118	0,012	− 0,179	0,013	− 0,031	0,023
FO	− 0,047	0,014	− 0,099	0,015	0,006	0,019	− 0,086	0,014	− 0,146	0,015	0,027	0,018
FR	0,086	0,013	0,098	0,014	0,120	0,015	0,090	0,014	0,137	0,015	0,063	0,015
GE	− 0,038	0,012	− 0,071	0,013	− 0,002	0,019	− 0,030	0,013	− 0,059	0,014	0,044	0,017

Table 4 Estimates of provincial grading standards by model across subjects 2/3

	Language						Mathematics					
	Model A		Model B		Model C		Model A		Model B		Model C	
Prov	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*
GO	0,053	0,021	− 0,017	0,023	0,112	0,025	− 0,067	0,022	− 0,154	0,023	0,081	0,025
GR	− 0,070	0,017	− 0,085	0,018	− 0,035	0,020	− 0,108	0,017	− 0,123	0,019	− 0,065	0,020
BZ	0,096	0,032	0,090	0,034	0,130	0,037	− 0,114	0,033	− 0,187	0,035	0,014	0,038
IM	− 0,055	0,018	− 0,096	0,019	− 0,006	0,022	0,021	0,018	0,005	0,019	0,070	0,021
IS	0,243	0,024	0,240	0,026	0,268	0,027	0,187	0,025	0,213	0,027	0,182	0,027
KR	0,044	0,017	0,134	0,018	0,046	0,017	0,027	0,017	0,124	0,019	− 0,103	0,018
LC	− 0,096	0,014	− 0,173	0,015	− 0,022	0,019	− 0,154	0,014	− 0,286	0,015	0,054	0,018
LE	0,113	0,011	0,114	0,012	0,161	0,013	0,108	0,012	0,125	0,013	0,103	0,013
LI	− 0,022	0,016	− 0,054	0,017	0,021	0,018	− 0,022	0,016	− 0,082	0,018	0,041	0,018
LO	− 0,097	0,016	− 0,154	0,017	− 0,036	0,021	− 0,008	0,016	− 0,059	0,018	0,095	0,020
LT	0,031	0,013	0,041	0,014	0,059	0,015	0,036	0,013	0,079	0,014	− 0,014	0,015
LU	− 0,138	0,016	− 0,155	0,017	− 0,096	0,019	− 0,088	0,017	− 0,108	0,018	− 0,051	0,019
MC	− 0,128	0,014	− 0,183	0,015	− 0,072	0,019	− 0,081	0,015	− 0,139	0,016	0,018	0,018
ME	0,148	0,012	0,218	0,013	0,158	0,013	0,129	0,013	0,197	0,014	0,051	0,014
MI	− 0,056	0,010	− 0,111	0,011	− 0,007	0,018	− 0,015	0,010	− 0,081	0,011	0,100	0,015
MN	0,033	0,014	− 0,005	0,015	0,087	0,019	0,027	0,014	− 0,015	0,015	0,112	0,018
MO	− 0,021	0,012	− 0,070	0,013	0,024	0,019	− 0,028	0,012	− 0,093	0,013	0,088	0,017
MS	− 0,064	0,019	− 0,094	0,020	− 0,023	0,022	− 0,100	0,019	− 0,135	0,021	− 0,029	0,022
MT	0,067	0,016	0,078	0,017	0,093	0,018	0,039	0,017	0,066	0,018	0,006	0,018
NA	− 0,080	0,010	0,007	0,011	− 0,068	0,011	− 0,046	0,010	0,065	0,011	− 0,208	0,012
NO	− 0,071	0,014	− 0,122	0,015	− 0,022	0,019	− 0,017	0,015	− 0,050	0,016	0,051	0,018
NU	− 0,050	0,020	− 0,004	0,021	− 0,029	0,021	0,009	0,020	0,095	0,022	− 0,089	0,022
OR	0,041	0,021	0,045	0,023	0,070	0,022	0,091	0,022	0,171	0,023	− 0,005	0,023
PA	0,032	0,011	0,114	0,012	0,036	0,012	0,031	0,011	0,117	0,012	− 0,103	0,012
PC	− 0,110	0,015	− 0,150	0,016	− 0,076	0,023	− 0,125	0,016	− 0,180	0,017	− 0,023	0,022
PD	− 0,160	0,011	− 0,216	0,012	− 0,106	0,018	− 0,152	0,012	− 0,238	0,012	− 0,012	0,016

Table 4 (continued)

	Language						Mathematics					
	Model A		Model B		Model C		Model A		Model B		Model C	
Prov	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*
PE	0,103	0,014	0,086	0,015	0,144	0,019	0,090	0,014	0,101	0,016	0,132	0,017
PG	0,029	0,012	− 0,018	0,013	0,073	0,019	− 0,021	0,012	− 0,057	0,013	0,057	0,017
PI	− 0,118	0,015	− 0,166	0,016	− 0,083	0,019	− 0,113	0,015	− 0,193	0,017	− 0,015	0,018
PN	− 0,054	0,014	− 0,105	0,015	− 0,001	0,020	− 0,055	0,015	− 0,139	0,016	0,082	0,019
PO	− 0,159	0,016	− 0,175	0,017	− 0,117	0,020	− 0,182	0,016	− 0,239	0,018	− 0,112	0,021
PR	− 0,012	0,013	− 0,070	0,014	0,037	0,020	− 0,031	0,014	− 0,078	0,015	0,052	0,018
PS	0,054	0,013	0,013	0,014	0,108	0,018	− 0,014	0,014	− 0,063	0,015	0,091	0,017
PT	− 0,175	0,016	− 0,215	0,017	− 0,129	0,020	− 0,124	0,016	− 0,152	0,017	− 0,063	0,019

Table 5 Estimates of provincial grading standards by model across subjects 3/3

	Language						Mathematics					
	Model A		Model B		Model C		Model A		Model B		Model C	
Prov	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*
PV	− 0,192	0,013	− 0,245	0,014	− 0,139	0,019	− 0,122	0,013	− 0,159	0,014	− 0,032	0,017
PZ	0,005	0,014	0,030	0,015	0,041	0,015	− 0,036	0,015	− 0,022	0,016	− 0,039	0,016
RA	−0,066	0,013	− 0,114	0,014	− 0,012	0,019	− 0,100	0,014	− 0,177	0,015	0,040	0,018
RC	0,106	0,013	0,164	0,014	0,112	0,013	0,125	0,013	0,197	0,014	0,019	0,014
RE	0,011	0,012	− 0,040	0,013	0,059	0,019	− 0,006	0,013	− 0,063	0,014	0,094	0,017
RG	0,082	0,014	0,134	0,015	0,100	0,015	0,031	0,015	0,088	0,016	− 0,050	0,016
RI	0,006	0,021	− 0,007	0,022	0,011	0,024	− 0,034	0,021	− 0,018	0,023	− 0,066	0,024
RM	0,008	0,010	− 0,005	0,011	0,036	0,016	0,036	0,010	0,046	0,011	0,054	0,014
RN	0,049	0,014	− 0,026	0,015	0,105	0,021	0,016	0,014	− 0,040	0,015	0,120	0,018
RO	−0,069	0,016	− 0,091	0,017	− 0,018	0,019	− 0,044	0,017	− 0,069	0,018	0,024	0,019
SA	0,019	0,011	0,079	0,012	0,037	0,012	0,029	0,011	0,105	0,012	− 0,058	0,012
SI	−0,156	0,016	− 0,197	0,017	− 0,097	0,022	− 0,190	0,016	− 0,265	0,017	− 0,030	0,020
SO	−0,150	0,017	− 0,217	0,018	− 0,094	0,021	− 0,190	0,017	− 0,326	0,019	0,011	0,021
SP	−0,090	0,018	− 0,145	0,019	− 0,050	0,023	− 0,011	0,018	− 0,042	0,019	0,044	0,021
SR	0,147	0,014	0,199	0,015	0,160	0,014	0,150	0,014	0,202	0,015	0,062	0,015
SS	0,083	0,014	0,124	0,015	0,097	0,015	0,137	0,015	0,221	0,016	0,002	0,016
SV	0,057	0,016	0,034	0,017	0,102	0,021	0,041	0,016	0,022	0,018	0,117	0,020
TA	0,208	0,012	0,247	0,013	0,230	0,013	0,142	0,012	0,201	0,013	0,047	0,013
TE	0,014	0,015	0,007	0,016	0,048	0,017	− 0,001	0,015	0,024	0,016	− 0,032	0,017
TN	0,261	0,012	0,231	0,013	0,307	0,017	0,109	0,012	0,019	0,013	0,262	0,017
TO	−0,059	0,010	− 0,104	0,011	− 0,012	0,016	0,015	0,011	− 0,024	0,011	0,089	0,014
TP	0,210	0,013	0,287	0,014	0,215	0,014	0,116	0,014	0,166	0,015	0,049	0,015
TR	0,083	0,017	0,055	0,018	0,106	0,021	0,032	0,017	0,006	0,019	0,067	0,021
TS	−0,140	0,019	− 0,184	0,020	− 0,089	0,025	− 0,112	0,019	− 0,191	0,021	0,055	0,024
TV	−0,146	0,011	− 0,204	0,012	− 0,087	0,017	− 0,130	0,011	− 0,228	0,012	0,025	0,016
UD	−0,118	0,013	− 0,166	0,014	− 0,065	0,019	− 0,124	0,013	− 0,222	0,015	0,036	0,018

Table 5 (continued)

	Language						Mathematics					
	Model A		Model B		Model C		Model A		Model B		Model C	
Prov	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*	*Beta*	*se*
VA	−0,134	0,011	−0,195	0,012	−0,072	0,017	−0,076	0,012	−0,145	0,013	0,048	0,015
VB	0,143	0,020	0,105	0,021	0,195	0,023	0,122	0,020	0,078	0,022	0,198	0,022
VC	0,076	0,019	0,064	0,020	0,127	0,021	0,146	0,019	0,131	0,021	0,203	0,021
VE	−0,026	0,012	−0,062	0,013	0,024	0,016	−0,067	0,012	−0,115	0,013	0,019	0,015
VI	−0,120	0,011	−0,181	0,012	−0,059	0,017	−0,041	0,011	−0,122	0,012	0,096	0,016
VR	0,051	0,011	0,009	0,012	0,105	0,017	0,004	0,011	−0,055	0,012	0,111	0,016
VT	−0,057	0,015	−0,082	0,016	−0,015	0,020	0,075	0,016	0,067	0,017	0,131	0,019
VV	0,012	0,018	0,097	0,019	0,024	0,019	0,036	0,018	0,116	0,020	−0,052	0,020

Acknowledgements The authors would like to thank participants to the VIII Seminario "I dati INVALSI: uno strumento per la ricerca e la didattica" (INVALSI 2023) Conference in Rome for the useful comments.

Funding Open access funding provided by Università degli Studi di Milano within the CRUI-CARE Agreement. This research was supported by Fondazione Cariplo under the project EDUFAIR – Mapping and Tackling Unfair Educational Inequality (Project Code 2024-1575), led by Moris Triventi, and funded within the INEQUALITIES RESEARCH 2024.

Data availability The data underlying this article were provided by INVALSI under license or by permission. Data will be shared upon request to the corresponding author with the permission of INVALSI. INVALSI is the National Institute for the Evaluation of the Education System, a research institute with legal status under European public law. The authors have signed an agreement to access these administrative data. Regarding this, three pieces of information: (1) First, access to the data is relatively straightforward, requiring only the signing of an agreement: https://serviziostatistico.invalsi.it/en. (Second, the authors are committed to assisting with data access in case of any issues. The site where data access can be requested is in English: https://serviziostatistico.invalsi.it/en/catalogo-dati. https://osf.io/6zae8/.

References

Argentin, G., Tamanini, G.: Il nuovo esame e le opinion degli studenti. In C. Tamaniniand C. Tugnoli (Eds.), Gli esami di Stato in Trentino: Ricerca e Laboratori (pp. 41-72),IPRASE-Studie Ricerche (2001)

Aucejo, E., Coate, P., Fruehwirth, J.C., Kelly, S., Mozenter, Z.: Teacher effectiveness and classroom composition: understanding match effects in the classroom. Econ. J. **132**(648), 3047–3064 (2022). https://doi.org/10.1093/ej/ueac046

Batruch, A., Geven, S., Kessenich, E., van de Werfhorst, H.G.: Are tracking recommendations biased? A review of teachers' role in the creation of inequalities in tracking decisions. Teach. Teach. Educ. **123**, 103985 (2023). https://doi.org/10.1016/j.tate.2022.103985

Betts, J.R., Grogger, J.: The impact of grading standards on student achievement, educational attainment, and entry-level earnings. Econ. Educ. Rev. **22**(4), 343–352 (2003). https://doi.org/10.1016/S0272-7757(02)00059-6

Bolletta, R. Studio sperimentale sull'assegnazione dei punteggi nelle prove scritte.Istituto Nazionale per la Valutazione del Sistema dell'Istruzione, Osservatorio Nazionalesugli Esami di Stato (2001).

Bowers, A.J.: What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. Educ. Res. Eval. **17**(3), 141–159 (2011). https://doi.org/10.1080/13803611.2011.597112

Bradbury, B., Corak, M., Waldfogel, J., Washbrook, E.: Too many children left behind: The US achievement gap in comparative perspective. Russell Sage Foundation 2015.

Braga, M., Paccagnella, M., Pellizzari, M.: Evaluating students' evaluations of professors. Econ. Educ. Rev. **41**, 71–88 (2014). https://doi.org/10.1016/j.econedurev.2014.04.002

Brookhart, S.M.: Graded achievement, tested achievement, and validity. Educ. Assess. **20**(4), 268–296 (2015). https://doi.org/10.1080/10627197.2015.1093928

Brookhart, S.M.: Teachers' grading practices: meaning and values. J. Educ. Meas. **30**(2), 123–142 (1993). https://doi.org/10.1111/j.1745-3984.1993.tb01070.x

Brunello, G., Cappellari, L.: The labour market effects of alma mater: evidence from Italy. Econ. Educ. Rev. **27**(5), 564–574 (2008). https://doi.org/10.1016/j.econedurev.2007.05.004

Carlana, M.: Implicit stereotypes: evidence from teachers' gender bias*. Q. J. Econ. **134**(3), 1163–1224 (2019). https://doi.org/10.1093/qje/qjz008

Chetty, R., Friedman, J.N., Rockoff, J.E.: Measuring the impacts of teachers I: evaluating bias in teacher value-added estimates. Am. Econ. Rev. **104**(9), 2593–2632 (2014a). https://doi.org/10.1257/aer.104.9.2593

Chetty, R., Friedman, J.N., Rockoff, J.E.: Measuring the impacts of teachers II: teacher value-added and student outcomes in adulthood. Am. Econ. Rev. **104**(9), 2633–2679 (2014b). https://doi.org/10.1257/aer.104.9.2633
Cinelli, C., Forney, A., Pearl, J.: A crash course in good and bad controls. Sociol. Methods Res. (2022). https://doi.org/10.1177/00491241221099552
Contini, D., Triventi, M.: Between formal openness and stratification in secondary education: Implications for social inequalities in Italy, in H.P. Blossfeld, S. Buchholz, J. Skopek & M. Triventi (Eds.). Models of Secondary Education and Social Inequality: An International Comparison, Cheltenham, UK: Edward Elgar, pp. 305-322 (2016).
Figlio, D.N., Lucas, M.E.: Do high grading standards affect student performance? J. Public Econ. **88**(9), 1815–1834 (2004). https://doi.org/10.1016/S0047-2727(03)00039-2
Fleischmann, M., Hübner, N., Marsh, H.W., Trautwein, U., Nagengast, B.: Investigating the association between the Big Fish Little Pond Effect and grading on a curve: a large-scale quasi-experimental study. Int. J. Educ. Res. **110**, 101853 (2021). https://doi.org/10.1016/j.ijer.2021.101853
Gay, G., Triventi, M.: Voti in italiano e competenze in lettura: come variano gli standard valutativi in Italia?, In U.S.R. (Ed.), Le competenze degli studenti lombardi. Il rapporto OCSE-PISA 2009 in Lombardia: risultati ed approfondimenti tematici (pp. 143-165), Brescia: Vannini (2011).
Gentrup, S., Lorenz, G., Kristen, C., Kogan, I.: Self-fulfilling prophecies in the classroom: teacher expectations, teacher feedback and student achievement. Learn. Instr. **66**, 101296 (2020). https://doi.org/10.1016/j.learninstruc.2019.101296
Heckman, J.: Policies to foster human capital. Nat. Bur. Econ. Res. **54**(1), 3–56 (1999). https://doi.org/10.3386/w7288
Hedges, L.V., Olkin, I.: Statistical Methods for Meta-Analysis. Elsevier Science, Saint Louis (2014)
Iacus, S.M., Porro, G.: Teachers' evaluations and students' achievement: a "deviation from the reference" analysis. Educ. Econ. **19**(2), 139–159 (2011). https://doi.org/10.1080/09645290903105277
Ichino, A., Polo, M., Rettore, E.: Are judges biased by labor market conditions? Eur. Econ. Rev. **47**(5), 913–944 (2003). https://doi.org/10.1016/S0014-2921(02)00269-6
Jan Skopek, Giampiero Passaretta, Socioeconomic Inequality in Children's Achievement from Infancy to Adolescence: The Case of Germany, Social Forces, Volume 100, Issue 1, September 2021, Pages 86–112, https://doi.org/10.1093/sf/soaa093
Jerrim, J., Vignoles, A.: Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes. J. R. Stat. Soc. Stat. Soc. **176**(4), 887–906 (2013). https://doi.org/10.1111/j.1467-985X.2012.01072.x
Lievore, I., Fedeli, E., Triventi, M.: The stricter the better? The impact of early teacher grading standards on students' competences development and academic track enrollment. Soc. Sci. Res. **124**, 103085 (2024). https://doi.org/10.1016/j.ssresearch.2024.103085
Lievore, Ilaria, Emanuele Fedeli, Moris Triventi.: 'The Stricter the better? The impact of early teacher grading standards on students' competences development and academic track enrollment' (2023)
Lievore, I., Triventi, M.: Do teacher and classroom characteristics affect the way in which girls and boys are graded?: a multilevel analysis of student-teacher matched data. Br. J. Sociol. Educ. **44**, 1–26 (2022). https://doi.org/10.1080/01425692.2022.2122942
Malouff, J.M., Thorsteinsson, E.B.: Bias in grading: a meta-analysis of experimental research findings. Aust. J. Educ. **60**(3), 245–256 (2016). https://doi.org/10.1177/0004944116664618
Modica, Salvatore, Valentino Dardanoni, Aline Pennsi.: 'Grading across schools'. B.E. J. Econ. Anal.; Polic. 9 (2009)
OECD: *Grade Expectations: How Marks and Education Policies Shape Students' Ambitions*. OECD. Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*, 2nd edn. Cambridge Univ. Press, Cambridge (2012)
Pearl, Judea. Causality: Models, Reasoning, and Inference. 2. ed. Cambridge: Cambridge Univ. Press. (2009)
Rosenthal, R., Jacobson, L.: Pygmalion in the classroom. Urban Rev. **3**(1), 16–20 (1968). https://doi.org/10.1007/BF02322211
Sestito, P., Tonello, M.: I differenziali nella qualità degli iscritti alle UniversitàItaliane: il caso delle Facoltà di Medicina e Chirurgia. Questioni di Economia e Finanza(Occasional Papers), Banca d'Italia (2011).
Sun, Y., Cheng, L.: Teachers' grading practices: meaning and values assigned. Assess. Educ. Princ. Polic. Pr. **21**, 326–343 (2014). https://doi.org/10.1080/0969594X.2013.768207
Triventi, M., Argentin, G.: The north-south divide in school grading standards: new evidence from national assessments of the Italian student population. Ital. J. Sociol. Educ. **7**(2), 157–185 (2015)
Tyner, A., Gershenson, S.: Conceptualizing grade inflation. Econ. Educ. Rev. **78**, 102037 (2020). https://doi.org/10.1016/j.econedurev.2020.102037

Westphal, A., Becker, M., Vock, M., Maaz, K., Neumann, M., McElvany, N.: The link between teacher-assigned grades and classroom socioeconomic composition: the role of classroom behavior, motivation, and teacher characteristics. Contemp. Educ. Psychol. **46**, 218–227 (2016). https://doi.org/10.1016/j.cedpsych.2016.06.004

Zanga, G., De Gioannis, E.: Discrimination in grading: a scoping review of studies on teachers' discrimination in school. Stud. Educ. Eval. **78**, 101284 (2023). https://doi.org/10.1016/j.stueduc.2023.101284

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.