THE STRICTER THE BETTER? THE IMPACT OF EARLY TEACHER GRADING STANDARDS ON STUDENTS' COMPETENCES DEVELOPMENT AND ACADEMIC TRACK ENROLLMENT

Ilaria Lievore*[1], Emanuele Fedeli[1], Moris Triventi[1]

* Corresponding author: ilaria.lievore@unitn.it

[1] University of Trento, Department of Sociology and Social Research, Via Verdi 26, Trento, 38122, Italy.

March 17th, 2023

Abstract

Despite the growing attention on teachers' grading practices in educational research, less attention has been dedicated to the consequences of teachers' grading standards on students' educational outcomes, especially in early stages of their scholastic career. This paper aims at filling this gap, analyzing the impact of teacher's severity in grading on students' competences development and academic track enrollment, and how it varies according to students' gender, socio-economic background and immigrant status. The analysis relies on Italian INVALSI-SNV data: information on 5th graders and their teachers are linked, and pupils are followed up to 8th and 10th grade, in which their competences and school track are recorded. Trough 2SLS regressions we demonstrate that being exposed to stricter grading in 5th grade leads to higher students' competences later on, and to higher probability to enroll in the most prestigious academic track, with no notable heterogeneous effects across students with different sociodemographic characteristics.

Keywords: teachers, grading standards, academic outcomes, student competences, school track enrollment.

1. Introduction

The analysis of grading practices, that is the way in which teachers grade their students, is at the core of an extensive literature in educational studies (for a review on teacher judgments, see Urhahne & Wijnia 2021). Grading practices have been shown to have a substantial impact on students' educational outcomes. Existing studies on this topic agree on the importance of grades in contemporary educational systems as well as in the labor market (Tyner & Gershenson 2020). Teacher grades can affect students' learning processes and how they perceive themselves in terms of ability and competence, which in turn have long-run implications for a number of students' life outcomes.

Among the immediate consequences of grading practices, there are students' placement in classroom, grade promotion and attendance habits (Bonner & Chen 2019; Gershenson 2016). Medium and long-run consequences might involve students' school choices, occupational decisions and earnings in adulthood (Borghans et al. 2016; Chetty et al. 2014; Bonner & Chen 2019).

Among educational institutions worldwide, grades serve as fundamental sorting and signaling mechanisms (Chowdhury 2018). However, these signals are not provided only to the students, who may need them in order to form an idea about their intellectual ability and, consequently, their possible educational future, but are captured and reproduced by many other players in the educational arena. With the increasing complexity of the educational systems, the significance of grades has assumed numerous facets, as many as the actors involved such as parents, teachers, principals, colleges and firms. For example, grades are important signals allowing a communication of students' academic achievement between schools and families. Parents use teachers' evaluations to make educational choices for their children and to efficiently track them in the school

system, and also to understand if their children need educational support (Jalava et al. 2015).

Correa and Gruver (1987) conceptualize grades as a fundamental parameter in the students' utility function. Since students care about teacher's perception of their achievement, students' effort and achievement may be affected by how teacher decide to grade students (Iacus & Porro 2008). Indeed, teachers can decide to adopt certain grading standards, that is the ability level needed by students in order to get a specific grade, or, in other words, how stringently teachers assess their students. Teachers with higher/harder grading standards tend to give good grades only to very high achieving students, who show very high competences and ability levels, while teachers with lower grading standards are likely to give good grades also to those students with average levels of ability, shrinking the grading scale.

Despite the large public debate on teachers' adoption and implementation of specific grading practices, especially in primary education, little empirical research focuses on how teacher can influence students' effort and motivation adopting specific grading standards, and on the associated educational consequences. On one side, students whose teacher adopts higher grading standards are those who need to put more effort and to study more if they want to achieve a good grade, and as a consequence, students might benefit in the long run in terms of competences (Iacus & Porro 2008). On the other side, higher grading standards may discourage students if the level of ability needed for achieving a good grade is too high. Moreover, it has been hypothesized that high grading standards may have heterogeneous effects among students (Betts & Grogger 2003) since motivation may be triggered differently according to students' characteristics (Becker & Rosen 1992) such as gender, socio-economic background and immigrant status.

The aim of this paper is to contribute to the empirical research on the effect of grading standards adopted in primary schools on educational outcomes, relying on a causal approach. The goal is to understand the effect of teacher grading standards measured in 5th grade on students' competences in 8th and 10th grade in two subjects – Language and Mathematics, and on school track in 10th grade. Additionally, the aim is to analyze whether teacher grading standards may have heterogeneous effects according to students' sociodemographic characteristics such as gender, socioeconomic background and migratory background.

The focus is on the Italian educational system, which is well suited for the study of teacher grading standards and their consequences, because teachers have a great deal of autonomy and independence in deciding their own grading practices, also when considering the school administration (Bracci 2009). On the one hand, this allows a certain degree of variation in grading practices already in early stages of educational career. On the other hand, Italian teachers' autonomy in deciding grading practices permits to explore the consequences of grading standards measured at the classroom level instead of at the school level, leading to more fine-grained results.

We rely on the INVALSI-SNV data, focusing on a cohort of Italian 5th grade students in the academic year 2013-14, which is followed up to 8th grade (a.y. 2016-17) and to 10th grade (a.y. 2018-19). This dataset allows to match students with their teachers and therefore to control for students', teachers' and classrooms' characteristics. Moreover, the availability of both teacher grades and students results in standardized tests allows us to create a measure of teacher grading practices.

The contribution of this article to the understudied topic of grading standards are threefold. First, as abovementioned, the Italian data used permits to explore the

consequences of grading standards at the classroom level, instead that at the school level, thus providing a more realistic and fine-grained perspective. Second, very few empirical research investigates the impact of more rigorous grading standards measured at the early stages of educational career (see Figlio & Lucas 2004 for an example), when there may be stronger effects on children self-perception of their ability, motivation and future educational choices (Facchinello 2020). Third, this contribution allows to causally assess the long-term consequences of having a teacher with high/low grading standards, not only in terms of competences but also in terms of academic track enrollment, which is a key transition point in the educational system associated with higher changes of later educational and labor market success (Barone et al. 2021; Triventi et al. 2021). This approach, combined with the analysis of heterogeneous effects on students with different characteristics, may have important implications also in terms of policy making, as discussed in the conclusions.

2. Literature Review on Grading Standards

In education, teachers' grading standards reflect the ability level needed by students in order to get a given grade. A teacher or a school with high grading standards tends to give good grades only to very high achievement, or to students who demonstrate very high levels of ability (Bonesrønning 2004). When measuring teacher grading standards with observational data (administrative or survey data), two pieces of information are usually needed at a classroom or school level: first, student ability measured by standardized test scores; second, student achievement measured by teacher assessment. The difference

between these two variables provides an idea about how stringently teachers assess their students compared to their actual competences.

Previous research about grading practices is rooted in the student-teacher interaction model proposed by Correa and Gruver (1987). In their utility-function of students, grades are thought as the product between student actual ability and teacher grading practices. But teachers can intervene in the relationship between students' competencies and assessed achievement through their grading practices, for example emphasizing the effort-component when formulating their judgements. In other words, through the practice of grading, teachers can both evaluate the sheer quality of students' work and at the same time motivate and encourage them to study (Walvoord & Anderson 1998).

The majority of empirical studies that proposed and supported the idea that grades – and the practice of grading – can have a significant impact on students' academic outcomes focused mostly on higher education and on college courses choice (Clark 1969; Gold et al. 1971; Hales et al. 1971; Cherry & Ellis 2005). However, relatively few authors have focused on the consequences of grading standards in early stages of the educational career. Concerning students' competences, Betts (1997; 1998) hypothesizes that more stringent grading standards will increase effort among students, and therefore their subsequent achievement. The author focuses on 7th and 10th grade students, and findings suggest that grading standards are important determinants of high school students' competences. Betts and Grogger (2003), analyzing 1,000 high schools, also find a positive effect of harder grading standards on students' performance in 12th grade, especially in the upper end of the grades distribution. Figlio and Lucas (2004) analyze the teacher-level grading standards on elementary students' achievement in Florida, using data on 3rd, 4th

and 5th grade. They find that higher grading standards seem to benefit students in language and mathematics test scores over time. On the contrary, Montmarquette and Mahseredjian (1989) analyze the effect of hard grading – grades set below the real achievement – on Canadian primary school pupils and found that they have a negative effect on test scores in Language, while they have no effect on test score in Mathematics. Some studies on Norway find that lower secondary school students who are exposed to harder grading standards perform better in mathematics (Bonesrønning 1999; 2004). The same results are confirmed by the study conducted by Iacus & Porro (2008) on a local sample of 20 lower secondary schools in Lombardy, an Italian region, in three subjects (language, science and mathematics). Concerning the impact of grading standards on students' educational choices at earlier educational stages, empirical research is even scarcer. To the best of the author's knowledge, only Betts and Grogger (2003) showed that harder grading standards have no significant effect on high school decision or college admission in the United States for the period 1989 to 1991.

The basic mechanism behind the effect of grading standards is thought to be related to their influence on students' motivation and effort (Iacus & Porro 2008). In this regard, some studies focused on how students' effort respond to being graded and ranked (Levitt et al. 2012; Jalava et al. 2015). On one side, setting higher grading standards may induce students to study more and to put more effort in order to satisfy the requirements imposed by their teachers. Indeed, when teachers have high grading standards, students need to increase their effort in studying if they want to achieve a good grade. On the other side, standards that are too high to reach can induce students to give up, and therefore they may have a detrimental effect on their competences, making the relationship between the two possibly non-linear. Facchinello (2020) found that even being graded instead of

not-being graded in the early stages of schooling have negative effects in effort among low-ability and low-SES students, who show lower motivation also later on.

It must be underlined that effort and motivation may be triggered differently according to students' ability and classroom composition. High grading standards can be more effective for already high achievers, because they have the cognitive resources to meet such high standards, but at the same time they may have a detrimental effect on less able students who tend to give up when they perceived standards are impossible to reach (Betts & Grogger 2003). Since high grading standards appear to have noticeable effects on students' competences who are already in higher position of achievement distribution, they may exacerbate achievement dispersion among students. However, rather than being detrimental for low-achieving students, higher grading standards may also translate in a smaller but still positive effect on their subsequent academic performance (Betts & Grogger 2003). The composition of the classroom can also act as a moderator of the impact of grading standards: indeed, in the United States high standards appear to be beneficial for high-achieving students when they are in low-achieving classes and for low-achieving students in high-achieving classes (Figlio & Lucas 2004).

It is important to note that given that academic performance is related to students' socio-demographic characteristics such as gender, ethnic background and social origin (Hattie 2008), more rigorous grading standards can also affect social inequalities in student achievement. Moreover, the response to grading incentives of different groups of students may not be uniform (Chulkov 2006). Yet, there is little evidence showing how students with different sociodemographic characteristics respond to the same grading standards. Concerning students' gender, Fallan and Opstad (2012), analyzing a sample of business school students, found that male students are more responsive to harder grading

practices, and they are more willing to put more effort if a change in grading standards requires more work in order to get an expected grade, while female students are less sensitive to change in grading standards. Boys are also more responsive than girls to short-term incentives, while girls are more intrinsically motivated (Vecchione et al. 2014). Motivation incentives may also trigger differently students with different socioeconomic and migratory background. For example, a recent paper investigating a sample of Italian students demonstrates that lower socioeconomic background is associated with lower level of intrinsic motivation and higher level of amotivation (Manganelli et al. 2021), and this may be associated with a more positive response to harder grading standards considering grades are a tangible, short-term reward. Other studies found that ethnic minority students show higher intrinsic motivation than native students, possibly to face their stigma awareness (Eccles et al. 2006; Gillen-O' Neel et al. 2011), therefore they may be less responsive to harder grading standards as a tool for manipulating effort. However, this pattern is not corroborated by a study on the Italian case, where children with immigrant parents display instead higher levels of extrinsic motivation than natives (Triventi 2020).

3. The Italian Grading System

In primary and secondary Italian schools, the Italian Ministry of Education (MIUR - *Ministero dell'Istruzione e del Merito*) offers indications about how teachers are supposed to grade their students. Teacher grades are assigned on a scale that goes from 1 to 10, where 6 is considered as the passing grade[1]. There are mainly two moments in which

[1] This is true considering the academic years under examination in this paper. However, a recent reform in Italy (2021) has introduced a new grading system in primary schools, that consists in

students and families meet with teachers and schools in order to know about the children's academic situation, and they correspond to two report cards. The first one is around February (first semester) and the second and definitive for that academic year is around June (second semester). If students report a grade below 6 in any subject at the end of the school year, they have to take an exam in that subject before the beginning of the new school year in September. If the result of such exam (*esame di riparazione*) is still insufficient, the student has to repeat the previous grade. Moreover, if students have three or more subjects with a grade below 6 in the final school report, they have to repeat the year, depending on the judgments of all professors for that students who join in order to decide case by case (*consiglio di classe*).

The report card usually shows average grades for each subject of all the examination undertaken by students until the end of the semester. The type of exams depends on the subjects, on the school regulation, but mostly on professors, who have a great deal of autonomy in deciding the exam structure (e.g., multiple choice questions vs open ended questions, oral exams vs written exams), the frequency for the evaluations as well as the grading criteria. Even if the MIUR offers some guidelines about grading practices, it is not known or clear the extent to which schools and teachers follow such guidelines: teachers usually decide their own grading criteria and grading practices, mostly according to each school's specific regulations.

After 8th grade, Italian students make their first educational choice concerning high schools. Interestingly, neither teacher grades nor teacher recommendations are

eliminating numerical grades in favour of more descriptive students' evaluation. This evaluation should reflect four levels of learning, approximately defined as "advanced", "intermediate", "basic" and "in the process of acquisition".

binding for entering specific tracks, and formally there are no access criteria. High school can be broadly divided in vocational schools (*istituti professionali*), technical schools (*istituti tecnici*) and lyceums (*licei*). Lyceums represent the academic track, and they can be further divided in traditional lyceums and other lyceums. Traditional lyceum includes the classical lyceum, focusing on humanities, and the scientific lyceum, focusing on math and science. Generally, this is considered the most prestigious and demanding track, that leads to university enrollment. Other lyceums are considered less prestigious, and include linguistic, socio-pedagogical and artistic lyceums. Technical and vocational schools, instead, usually lead to entering the job market. Despite Italian upper secondary education is strongly stratified, university enrollment is formally open, and it does not depend on previous academic performance, or final grade: the basic requirement is having a 5-year high school diploma, although access to some universities is regulated by admission tests.

With regard to grading practices and grading standards, the topic has attracted public attention especially for what concerns the North-South divide in upper secondary education. In this respect, Argentin and Triventi (2015) examined the geographical heterogeneity in grading standards in two subjects and across the three educational levels constituting compulsory education in Italy. The results indicate that southern regions are generally characterized by lower grading standards, meaning that teachers are more generous in assigning grades for a given level of competence, especially considering high performing students. Yet, the Italian context is characterized by high levels of heterogeneity, even among provinces or schools within the macro-areas.

4. Material and Methods

4.1 Data

The empirical analysis is based on data collected by INVALSI-SNV (Italian National Institute for the Evaluation of the Education System). The main aim of INVALSI is to perform periodic, systematic and standardized assessments on students' competences. The SNV (National Evaluation System) data contain socio-demographic variables for the whole population of students enrolled in specific grades and academic years. Additionally, they contain information on both teacher assessment of student achievement (teachers' grades) and student scores in standardized tests in Language and Mathematics (INVALSI test score[2]). Starting from the year 2012, INVALSI handed out for the first time a CAWI questionnaire addressed to a random sample of Language and Mathematics teachers for specific grades. The questionnaire collects information on both teachers' socio-demographic characteristics, teaching habits and practices.

The selected sample for our analysis includes the cohort of 5th grade students in 2013/14. Leveraging the availability of unique classrooms identifiers, we matched this dataset with information from their teachers sampled in the same academic year. Students are then followed through their academic career using a student unique identifier (the SIDI code). The cohort of 5th grade students is therefore followed over time, linking

[2] INVALSI test score results are also corrected in order to reduce the risk of cheating during test administration (INVALSI, 2018).

information in 8th grade in the academic year 2016/17 and in 10th grade in the academic year 2018/19[3].

Since the language test and the mathematics test are administered in different days, some students may have been absent on one of the two days. In order to compare results across subjects, the analysis rely on a unique analytical sample that includes the considered outcomes in both the two subjects, respectively in grade 5, 8 and 10. Our final sample includes 9,370 students[4].

4.2 Measuring Teacher Grading Standards

The main independent variable is teacher grading standards, measuring how stringent teachers evaluate their students, relatively to the student achievement measured through the INVALSI test score. Standardized test scores are designed to capture specific competences acquired by students during their educational career (Heckman & Kautz 2014) and are considered more objective than grades, also because they are usually blinded evaluated. Following Betts and Grogger (2003), a measure of teacher grading standards is construct using two pieces of information: students' grades in Language and Mathematics, as a measure of how a student stands relatively to their classmates, and students' test score in language and Mathematics, as a measure of the student competences relatively to all Italian students. Teacher grading standards are estimated for each class, therefore all the students in the same class have the same teacher grading

[3] This is the unique cohort that was possible to follow, since in 2019/20 the INVALSI test was not administered due to the COVID-19 pandemic.

[4] Analysis performed with samples including the higher number of cases as possible for language and mathematics, therefore different samples for the two subjects, lead to almost identical results.

standards. Relying on two different regressions for Mathematics and Language, grading standards estimates are obtained by regressing separately students' test score in mathematics and in language competences on students' GPA (grade point average) in mathematics (eq. 1a) and in language (eq. 1b) respectively, plus a vector of classroom dummies:

$$Test\ Score\ Maths_{ic} = \beta_1 Teacher\ Grade\ Maths_{ic} + \alpha_2 Classroom\ Dummies_c + \varepsilon_{ic} \quad (1a)$$

$$Test\ Score\ Lang_{ic} = \beta_1 Teacher\ Grade\ Lang_{ic} + \alpha_2 Classroom\ Dummies_c + \varepsilon_{ic} \quad (1b)$$

Coefficients of classroom dummies are the estimated grading standards in Language and Mathematics. This implies that if there is a variation across teachers, a class with higher α_c has higher/harder grading standards. If $\alpha_{c1} > \alpha_{c2}$, a student in class 1 is exposed to higher grading standards respect to a student in class 2: the two students have an equal GPA in subject *s*, but the student in class 1 has a higher test score in subject *s* than the student in class 2.

4.3 Outcome Variables and Control Variables

The goal of the analysis is estimating the effect of grading standards in the 5th grade (t = 0, primary education) on students' subject-specific competences when students are in 8th grade (t = 3, lower secondary education) and in 10th grade (t = 5, upper secondary education). Moreover, the aim is assessing the effects of such grading standards in primary education on the probability of being enrolled in traditional lyceum when students are in 10th grade (t = 5). To sum up, the outcome variables are: 1) student competences in Language and Mathematics in grade 8, 2) student competences in Language and Mathematics in grade 10; and 3) the probability of being enrolled in

traditional lyceum in grade 10. Competences are measured through the INVALSI test score, while the school track is retrieved from the administrative register.

The INVALSI-SVN data allows us to control for a rich set of variables that concern student characteristics and demographics. We selected control variables following general recommendations from the causal graph literature (e.g., Cinelli et al. 2002). Among these, a measure of students' achievement in t = -1, as self-reported average grade at the end of 4^{th} grade is included. It is reasonable to assume that this measure captures what could have influenced parents' educational choices up to t = 0. Regarding teachers, control variables include a set of demographics (age, gender, educational credentials, parental education) together with some indicators associated with teacher effectiveness, such as type of contract and teaching to test information. At the classroom level, control variables include share of females, share of students with high socioeconomic background, share of immigrant students and class size. For a more detailed description of the control variables, see appendix Table A1 and Table A2.

4.4 Methods

The goal of this study is to causally identify and estimate the average treatment effect of being exposed to a particular grading standard in 5^{th} grade on students' subsequent academic competences (in language and mathematics) and their school track placement in upper secondary education. In order to do so, we developed two distinct approaches, which rely on different assumptions. In the first approach, we provide an identification of the causal effect controlling for an extensive array of individual, teachers, and classroom characteristics, and introducing school fixed effects to control for unobserved

characteristics at the school level. This first approach relies on three main assumptions: 1) No reverse causality between treatment and outcomes; 2) No confounding bias (at the individual and higher levels); 3) Teachers' characteristics are good proxies of teacher proclivities. Given that treatment and outcomes are measured in distinct moments of time, and given that we control for previous academic competences, the first assumption is likely to be satisfied. To empirically support the plausibility of the second assumption, a randomization check is performed, to evaluate whether grading standard "predicts" invariant student characteristics (Pei et al. 2019)[5]. Results show appreciable as-good-as-random distribution of grading standards across students, with the exception of students' socioeconomic status (see Appendix Table A3 and A4).

However, the third assumption according to which teachers' characteristics are good proxies of teacher proclivities might be violated. Teacher proclivities may affect grading standards due to observed and unobserved student characteristics, and they may depend also on teacher-student interactions (Aucejo et al. 2022). To control for such bias, in our second approach, we aim to account for potential remaining unobserved heterogeneity by relying on an instrumental variable design, where the instrument is the grading standards of other classrooms in the same schools. Our intuition is to exploit a teacher peer effect[6] within the school, where teachers are likely to discuss and compare grading practices, therefore to influence each other's grading practices. The two

[5] We test for consistency with as-good-as-random assignment of treatment in order to assess whether our treatment is randomly distributed across student categories (e.g., based on gender, ethnic origin, socioeconomic status); A low degree of selection and a rich set of controls support the plausibility of the lack of relevant omitted variable bias.

[6] Reflection is not an issue as outlined by Hernán and Robins (2006). First, IV estimation does not rely on assumptions about the causal ordering between the instrument and the endogenous regressor. (Birkelund & van de Werfhorst 2022).

approaches have in common the use of school fixed effects, to control for heterogeneity of grading standards (Argentin & Triventi 2015). Their goal is estimating the total effect of grading standards, therefore post-treatment variables which might lead to a bias are not included in the regressions (Elwert & Winship 2014). The estimation strategies follow the two approaches. In the first approach, three linear OLS regressions are estimated[7], with the following general specification:

$$Outcome = \beta_0 + \beta_1 \hat{\alpha}_{c_{t0}} + \beta_2 X_{i_{t0}} + \beta_3 T_{c_{t0}} + \beta_4 Z_{c_{t0}} + \mu_{s_{t0}} + \varepsilon_{ic} \quad (2)$$

The three outcomes are: 1) Mathematics and 2) Language competences measured three and five years after the 5th grade, when a new sorting of students in the lower and upper secondary education occurred; 3) probability of being enrolled in a traditional lyceum five years later. In the equation, $\hat{\alpha}_{c_{t0}}$ is the treatment of interest measuring teachers' grading standards; $X_{i_{t0}}$ is a vector of individual characteristics; $T_{c_{t0}}$ is a vector of teacher characteristics; $Z_{c_{t0}}$ is a vector of classroom characteristics and $\mu_{s_{t0}}$ are school fixed effects. In the second approach, the previous equations are modified by including the first stage of a 2LS estimation:

$$GS_s = \beta_0 + \mu_{s_{t0}} + \beta_1 \hat{\alpha}_{other\ class_{t0}} + \beta_2 X_{i_{t0}} + \beta_3 T_{c_{t0}} + \beta_4 Z_{c_{t0}} + \varepsilon_{ic} \quad (3)$$

[7] In order to check for the nonlinearity of the relationship, analyses are performed on the same models adding a quadratic term to the treatment variable. However, Likelihood-ratio test, AIC and BIC show no differences between the linear regression and the quadratic regression when including the control variables, even when the quadratic term is statistically significant. Results for model 3 are shown in appendix Table A5, A6, A7 and Figures A1, A2. The linearity of the relationship may be due to the fact that grading standards are measured in primary schools, where grading standards may generally be not particularly heterogeneous and overall not particularly severe.

Where GS_s represents the subject-specific grading standards as estimated in equation 1, $\hat{\alpha}_{other\ class_{t0}}$ stands for the grading standards adopted in the other classrooms in the school (the instrumental variable); all other terms are previously defined.

An additional empirical issue we tackled in the estimation of our statistical models refer to longitudinal missing values, commonly known as 'panel attrition'. Indeed, following students through their academic career implies an attrition that causes a significant loss of cases from the initial sample. This is due to several factors, such as grade retention, students transferring, non-reporting of SIDI codes by school administrations and also potential misclassification of SIDI codes. This may lead to a possible selection of high performing students that may in turn affect the estimates. We take into account the possible selectivity of students observed throughout the entire time span considered (from 5th to 10th grade), by correcting the estimates with an inverse probability weighting (IPW) approach, which has been shown to be effective a wide range of settings (Seaman & White 2013). In order to construct IPWs, we estimated a binomial logistic regression on the probability of being observed in the 10th grade among 5th grade students with valid information, as a function of a number of students' characteristics[8]. Then we computed predicted probabilities based from this model, we created weights as the inverse of the predicted probability and incorporated the regression estimations.

[8] The covariates are: gender, quarter of birth, ethnic background, regularities of studies, geographical area, attendance to infant school, attendance to kindergarten, socioeconomic background, INVALSI test score in Mathematics and Language, test anxiety. In order to control for the validity of IPWs, we perform additional analyses with weights attributed to students as the mean of the respective quantile of IPW, and results show no significant differences.

5. Results

5.1 Descriptive Analysis of Grading Standards

Figure 1 represents the distribution of grading standards (standardized) in the two subjects. In order to understand how grading standards are interpreted, it is important to recall that, through the analysis, the measure of how stringent the teacher is when assigning grades is not interpretable in absolute terms. Indeed, the construction of GS is relative to the selected sample – therefore to the selected teachers: the estimated effect on students' educational outcomes is interpretable as a change in severity within the selected population.

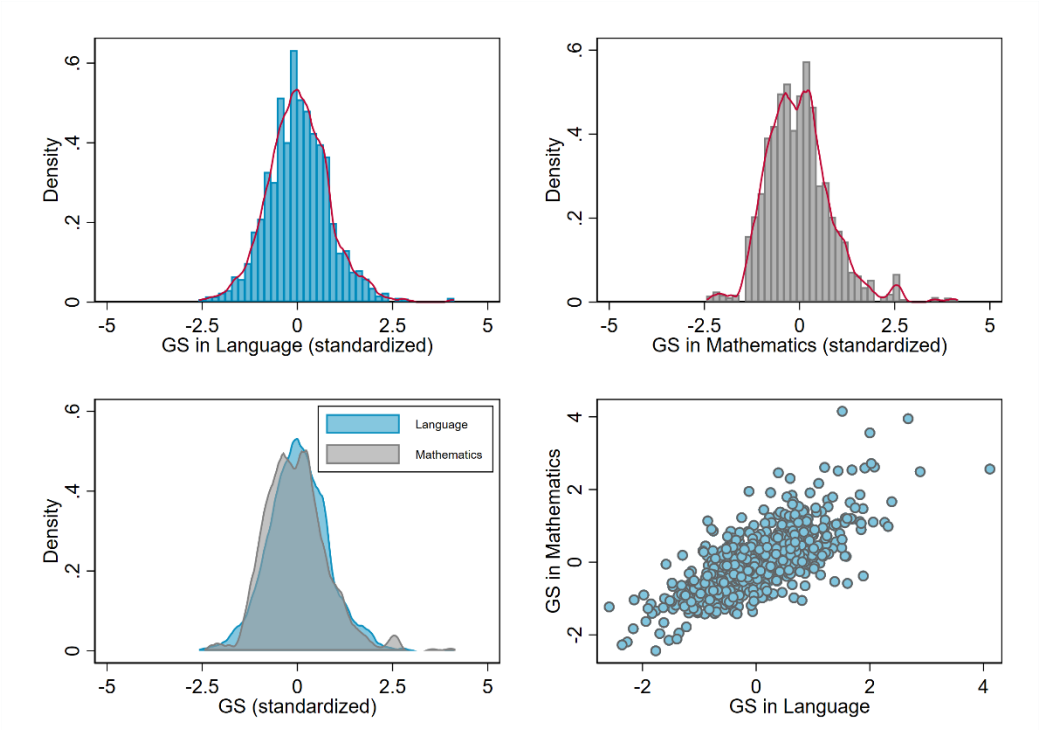

Figure 1*:* Distribution of grading standards in Language and Mathematics (N = 9,370) and correlation between grading standards in Language and grading standards in Mathematics.

However, considering that the analysis relies on a random sample of the whole population of Italian students in 5th grade in the academic year 2013-14, it is reasonable to assume that grading standards manages to virtually capture the whole spectrum of teacher severity in the considered grade.

In order to understand how to interpret grading standards in Language and Mathematics, it may be useful to rely on Table 1, in which classrooms with the lowest and the highest grading standards in Language are reported and compared with the classroom average score and the average grade. In order to facilitate the interpretation, we also report the mean for the eight values for GS, score and grade. It is noticeable how classes with lower grading standards, therefore having a more generous teacher, have poor INVALSI test score results compared to classes with higher grading standards, therefore having a stricter teacher (mean score of 179 against mean score of 262). At the same time, the average grade of classrooms with lowest grading standards is significantly higher than the one of classrooms with highest grading standards (9.1 against 7.8). These classrooms, with both lowest and highest GS, are the ones for which the distance between INVALSI score and grade is bigger: ideally, in a continuous that goes from the strictest teacher to the most generous teacher, it is possible to imagine a classroom for which the distance between INVALSI score and grade is null. The same identical patterns happen considering grading standards in Mathematics.

	Language			*Mathematics*		
	Grading Standard (std)	INVALSI score (classroom average)	Grade (classroom average)	Grading Standard (std)	INVALSI score (classroom average)	Grade (classroom average)
Classrooms with lowest GS						
	-2.59	173.9	9.1	-2.44	162	9.1
	-2.36	182.7	9.4	-2.28	172	9.5
	-2.27	172.7	8.8	-2.2	163.3	9.4
	-2.16	177	8.9	-2.15	145.4	8.1
	-2.15	182.2	9.4	-2.11	167	8.9
	-1.98	195.1	9.6	-1.97	158.9	8.3
	-1.94	186	9.2	-1.95	156.7	8.3
	-1.90	165.3	8.4	-1.83	173.6	9
Mean	**-2.17**	**179.4**	**9.1**	**-2.12**	**162.4**	**8.8**
Classrooms with highest GS						
	2.06	233.5	7.1	2.56	293.1	8.9
	2.08	258.9	8.3	2.59	272.9	7.8
	2.26	243.1	7.3	2.61	281.5	8.1
	2.32	249.9	7.6	2.62	287.8	8.3
	2.39	264.3	7.9	2.72	271.1	7.5
	2.67	259	7.4	3.56	296.9	7.6
	2.89	269.6	8.1	3.94	313.2	7.4
	4.11	317.9	8.9	4.15	316.3	7.7
Mean	**2.60**	**262**	**7.8**	**3.2**	**291.6**	**7.9**

Table 1*:* Bottom/top 8 classrooms with teachers having lower/higher GS in Language and Mathematics, and respective classroom average of INVALSI test score and classroom average grade (N = 9370).

Note: INVALSI test score and grade are shown in their original scale: INVALSI score has mean 200 and S.D. 40; grades are in a scale from 1 to 10.

5.2 Effect of GS on Student Competences

In this section we report our findings related to the effect of grading standards on student competences in subsequent educational levels. Figure 2 reports the average marginal effects of teacher grading standards in 5^{th} grade on INVALSI test score in 8^{th} grade and

in 10th grade in both Mathematics and Language, derived from four different models for each subject. The first model, which includes the treatment alone, shows that an increase of one standard deviation (SD, hereafter) in teacher grading standards corresponds to an increase of about 0.08 SDs in Language competences, both in 8th and 10th grade. For mathematics competences, one standard deviation in teacher grading standards is associated to a variation of 0.06 SDs in competences in grade 8 and nearly of 0.10 SDs in competences in grade 10.

Comparing the specification of model 1 to the specification of model 2, where students' demographic characteristics and previous ability are included, the coefficients increase for both subjects. In model 3 and 4, where fixed effects at the school level and the instrumental variable approach are adopted, we observe that an increase of one SD in teacher grading standards corresponds to an increase of about 0.15 SDs in students Language and Mathematics competences at the end of lower secondary education (grade 8), and in Mathematics competences in upper secondary education (grade 10). The increase in Language competences in grade 10 is slightly smaller, around 0.12 SDs.

In order to understand the magnitude of the increase in competences, results can be interpreted on the original scale of the INVALSI test score. The average result in INVALSI is around 200 points, with a standard deviation of 40. An increase of 1 standard deviation in grading standards correspond to an increase of about 6 points in the INVALSI test for both mathematics and language competences in 8th grade, and of about 5 to 6 points in 10th grade competences. All the model specifications suggest that 5th grade students who are exposed to a teacher with higher grading standards, or to a more severe teacher, are more likely to benefit in terms of competences gained three and five years later.

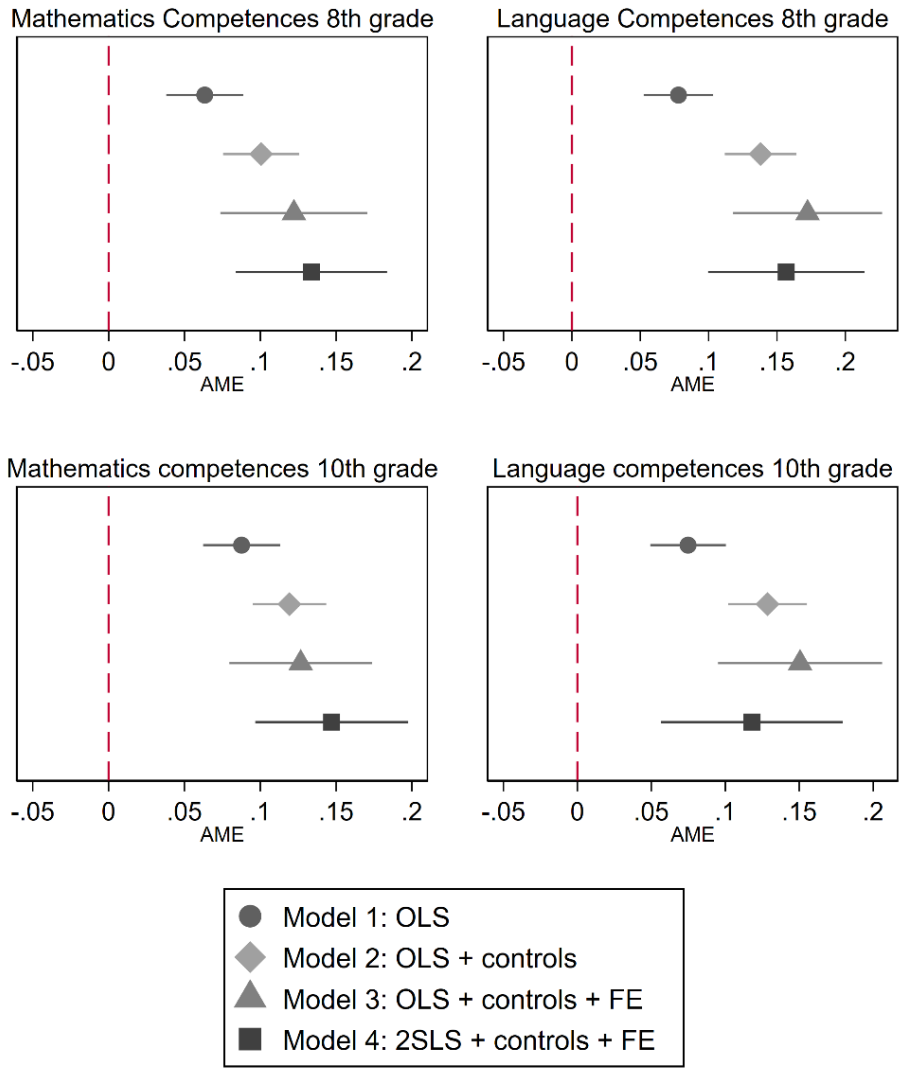

***Figure 2**:* Average marginal effects of GS in 5th grade on INVALSI test score in 8th and in 10th grade in Mathematics and Language competences; coefficients derived from OLS; N = 9370; 95% C.I.

Note: Model 1 controls for treatment. Model 2 includes students' sociodemographic and previous performance, teacher characteristics and classroom composition. Model 3 includes school fixed effect. Model 4 includes the instrumental variable.

The next goal is understanding whether such positive impact of having a stricter teacher is similar or equal for students with different socio-demographic characteristics, therefore coming from different socioeconomic background, with opposite gender or with a migratory background or not. Importantly, since in this analysis we adjust for teachers' grades in the 4th grade, what we are looking at is the possible heterogeneous reactions to being exposed to certain grading standards across categories of students identified by ascriptive characteristics but with comparable levels of previous academic performance. Figure 3 shows the average marginal effects of grading standards on students' competences measured later on in time, by students' migratory background, gender and socioeconomic background. Coefficients are derived from model 4, with all control variables, fixed effects at the school level and the IV specification.

Results show that the positive effect of grading standards on students' Language and Mathematics competences is pretty similar across students with different gender, migration background and social origin. The effect sizes are in most of the cases very similar and the 95% confidence intervals are widely overlapped. Concerning migratory background, instead, the inspection of effect sizes suggests a potential negative effect of high language grading standards for immigrant students in high school and a null effect for lower social background students. Unfortunately, the wide confidence intervals, make it difficult to provide a firm conclusion on these results based on our data.

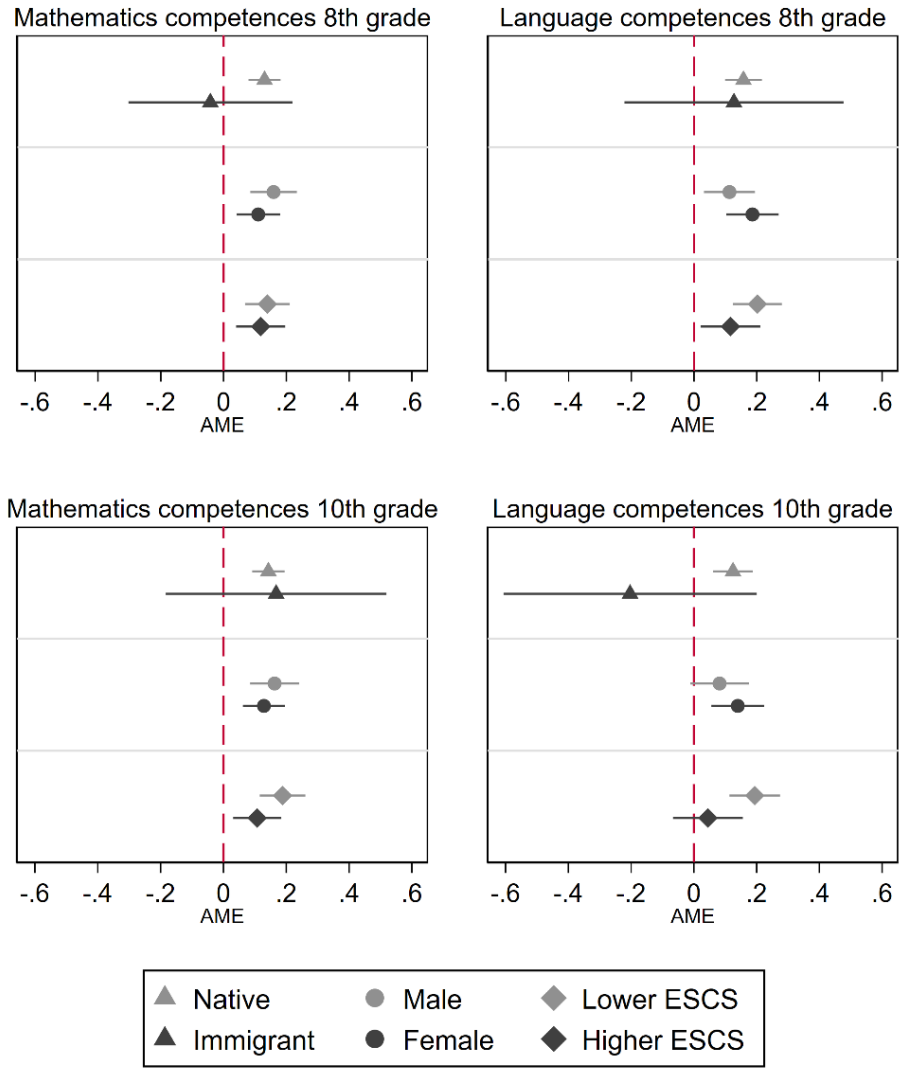

***Figure 3**:* Average marginal effects of GS in 5th grade on INVALSI test score in 8th and in 10th grade in Mathematics and Language competences by student characteristics: immigrant status, gender, ESCS; coefficients derived from OLS; N=9370; 95% C.I.

Note: Coefficients derived from model 4 (all control variables, fixed effect at the school level, iv specification)

5.3 Effect of GS on Student Probability of Being Enrolled in a Traditional Lyceum

In this section, the impact of teacher grading standards in 5th grade on students' probability of being enrolled in a traditional lyceum in 10th grade, rather than being enrolled in other lyceums, technical or vocational schools, is shown. In the analyzed sample, 38% of students are enrolled in traditional lyceums in grade 10. Figure 4 shows a positive effect of having a stricter teacher in 5th grade on the probability of being in a traditional lyceum rather than a non-traditional lyceum, or in a vocational or a technical school. In the baseline model specification without covariates (model 1), an increase of 1 standard deviation in teacher grading standards corresponds to an increase of 2 percentage points in the probability of being enrolled in lyceum having strict mathematics teacher, and of 4 percentage points in the probability of being enrolled in traditional lyceum having strict language teachers.

When including students' sociodemographic characteristics and previous ability, the effects slightly increase. There are no substantive differences between model 3 (with school fixed effects) and model 4 (iv specification). The effect of an increase of one standard deviation in the strictness of mathematics teacher in 5th grade corresponds to an increase of 5 percentage points in the probability of being enrolled in a traditional lyceum in 10th grade. Considering teacher grading strictness in language, the increase in the probability is 4 percentage points.

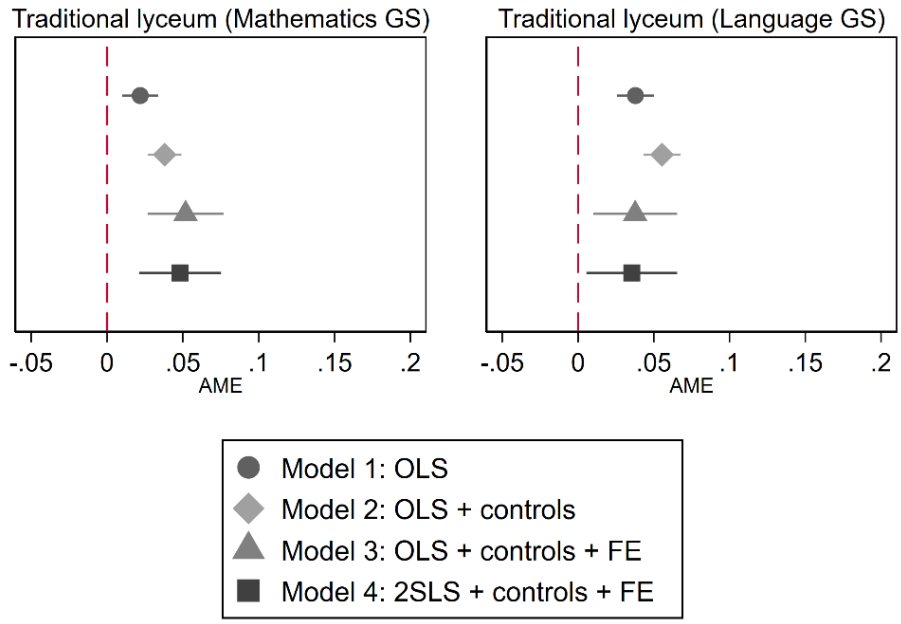

Figure 4*:* Average marginal effects of GS in 5th grade in Language and Mathematics on the probability of being enrolled in a traditional lyceum in 10th grade; coefficients derived from OLS; N = 9370; 95% C.I.

Note: Model 1 controls for treatment. Model 2 includes students' sociodemographic and previous performance, teacher characteristics and classroom composition. Model 3 includes school fixed effect. Model 4 includes the instrumental variable.

The investigation of heterogeneous effects for students with different migratory background, gender and socioeconomic background is presented in Figure 5. Results show that having a mathematics teacher with higher grading standards at the end of primary education has a positive effect on the chances of being enrolled in a traditional lyceum 5 years later, and this effect is similar across students with different sociodemographic characteristics, but comparable early academic performance. The exception are immigrant students, for which the coefficient is not statistically significant probably because of the low sample size. Results are more controversial when considering language teachers.

It appears that immigrant students may benefit more from having a strict teacher in language in 5th grade comparing to native students, for which the effect is close to zero. Female students do not benefit in terms of enrollment in traditional lyceum from having had a strict teacher in Language in 5th grade compared to male students. Finally, looking at heterogeneous effect of grading standards, results indicate that students' ESCS does not moderate the positive effect of Language teacher grading standards on the probability of being enrolled in a traditional lyceum.

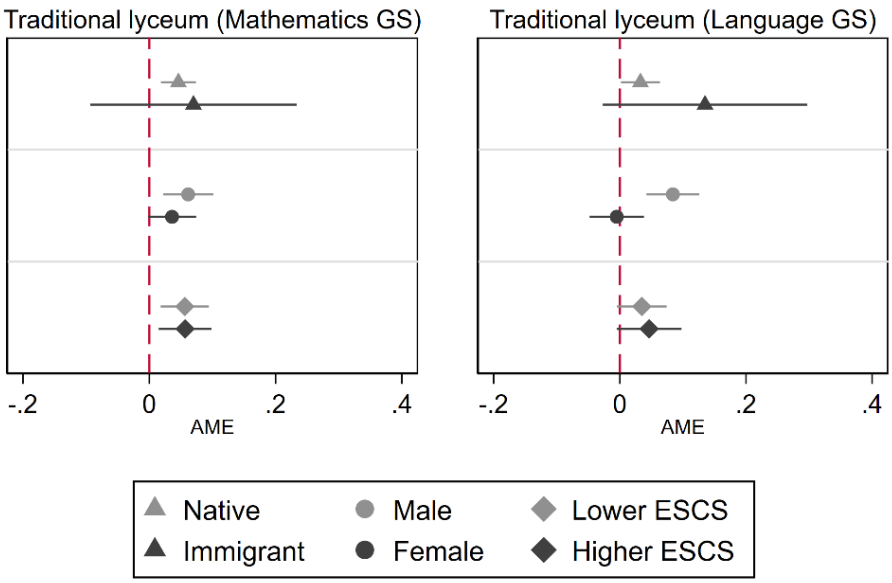

***Figure 1**:* Average marginal effects of GS in 5th grade in Language and Mathematics on the probability of being enrolled in a traditional lyceum in 10th grade by student characteristics: immigrant status, gender, ESCS; coefficients derived from OLS; N=9370; 95% C.I.

Note: Coefficients derived from model 4 (all control variables, fixed effect at the school level, iv specification)

6. Discussion

This paper addressed the issue of teacher grading standards in primary school, and how they affect important educational outcomes. The focus is on children's competences development and enrollment in academic tracks such as traditional lyceums in Italian schools. Grading standards is a measure reflecting of how strict the teacher is when evaluating and assigning grades to their students. Specifically, grading standards reflect the level of students' competences needed in order to get a specific grade, therefore students with similar competence but belonging to different classrooms may get higher grades when their teacher has lower grading standards and vice versa. Previous results suggest that through grading practices, and grading standards, teachers can manipulate students' effort and motivation: higher standards may induce students to increment their effort in order to satisfy teachers' requirements if they aspire to get a good grade, and, as a consequence, students can boost their competences development (Betts & Grogger 2003; Iacus & Porro 2008) and more generally they can benefit in terms of educational outcomes and choices. On the other hand, if teachers have grading standards that are too high to reach, it can induce students to give up, and this may have a detrimental effect on students' educational outcomes.

In line with most previous empirical findings (see Montmarquette and Mahseredjian 1989 for an exception), results show a positive effect of grading standards measured in primary school (5^{th} grade) on both subject specific competences and probability of being enrolled in a traditional lyceum in high school. Results hold considering competences in Language and Mathematics and looking at different time points – three and five years after the treatment.

When looking at heterogeneous effect, results are less clear-cut. Concerning students' competences development throughout the years, it seems that 1st and 2nd generation immigrant students benefit less than native students from having a strict teacher, in both Language and Mathematics. Even if it is difficult to interpret results based on the estimates because of the low sample size of immigrant students, they may suggest that the effect for immigrant students is nearly zero, or even negative considering Language competences measured in 10th grade. This may be partially explained by the struggle that especially 1st generation immigrant students face in learning a new language, and having a strict teacher in primary school in Language may discourage them from learning and studying the subject, leading to detrimental consequences for their competences later on in time. Concerning socioeconomic background, it seems that high ESCS students may benefit less from having a teacher with high grading standards in 5th grade, and the effect is null considering competences in Language measured in 10th grade. Focusing on the probability of being enrolled in a traditional lyceum, the positive effect of having a mathematics teacher with high grading standards in primary school is similar across students with different sociodemographic characteristics. Instead, the effect of having a strict language teacher is less straightforward: the effect is no longer significant looking at female students compared to male students, and looking only at students' socioeconomic background, but it becomes larger considering immigrant students compared to native students. Overall, despite such minor signs of heterogeneity on the basis of migration background, our main conclusion is that in the Italian context, higher grading standards seem to have positive or at best null impacts on a variety of students' outcomes in lower and upper secondary education. We did not detect clear evidence for

specific detrimental consequences for specific categories of students identified on the basis of their socio-demographic characteristics.

7. Conclusions

In line with previous results on the topic conducted mostly in the United States, this work suggests that stricter grades might be overall beneficial for students' subsequent educational outcomes in Italy, even when measured in primary school. Interestingly, empirical investigations focusing on grading practices in primary education are scarce, even if it is considered a crucial moment in students' educational journey in terms of competences development (Facchinello 2020). Indeed, adopting hard grading standards on 10 years old pupils may have strong implications that deserve particular attention. For instance, higher grading standards within a classroom imply increased inequalities among students, that can push pupils to benefit from an early categorization and ranking of their abilities in comparison with their peers. This may have positive consequences on their motivation, self-esteem, self-identity, as well as on their endurance and effort, and consequently on their educational competences and trajectories. It is important to underline that this may hold in the analyzed context, in which relatively harder grading standards are in absolute terms not particularly hard, considering that grades of 5^{th} grade pupils are overall high and teachers are in general pretty generous when attributing marks in this educational level.

This work underlines how teachers should be aware of how specific grading practices, and particularly those considered as severe ones, may actually help their students, independently of students' sociodemographic characteristics. Following the work of

Facchinello (2020), this paper suggests that the social scientists should dedicate more attention on the topic of the grading system, particularly in the early stages of the educational career, in order to investigate aspects that have been somehow overlooked by the educational literature and might have important implications also in terms of educational policy making. This is especially true in the Italian context, in which the 2021 reform on primary schools eliminated numerical grades and promoted descriptive students' evaluations. Grading practices may have important effects on how students perceive themselves and their ability. Indeed, if students receive inflated grades – higher than what they actually deserve – their parents and themselves may believe that they are prepared for specific situations (e.g., highly demanding academic education), while they are not. Moreover, if very skilled and prepared students get the same grades as their less-prepared colleagues, this might instill a sense of frustration and demotivation in the former, thereby leading to reduced effort in schooling and participation in classroom activities (Finefter-Rosenbluh & Levinson 2015). In the long run, the entire work-ethic of students can result deteriorated from this process, since it may suggest that hard work is not needed for achieving educational success (Chowdhury 2018). This study shows that a reform of the grading practices in elementary school has been implemented without a careful consideration of the pros and cons and without a full consideration of the actual grading practices adopted by Italian teachers. Our findings seem to suggest that in a context of overall generous evaluations towards children in primary education, adopting relatively stricter standards appear not to have negative consequences and, for most of students' categories, to positively affect their subsequent educational outcomes.

8. Acknowledgments

An earlier version of this paper was presented at the VI seminar INVALSI data (2021) and at the 5th Joint Interdisciplinary Graduate Conference (2022). The authors thank all the participants, and in particular Gianluca Argentin, Gabriele Ballarino and Dalit Contini, who gave us useful insights on these occasions. Usual disclaimers apply.

9. Funding

This work was supported by the Compagnia di San Paolo Foundation (Italy), through the international cooperation project INEQUALITREES, grant number A130902.

10. References

Argentin, G., Triventi, M. (2015): The North-South Divide in School Grading Standards: New Evidence from National Assessments of the Italian Student Population, in *Italian Journal of Sociology of Education*, 7(2):157-185.

Aucejo, E., Coate, P., Fruehwirth, J.C., Kelly, S., Mozenter, Z. (2022): Teacher Effectiveness and Classroom Composition: Understanding Match Effects in the Classroom, in *The Economic Journal*, 132(648):3047-3064.

Barone, C., Triventi, M., Facchini, M. (2021): Social Origins, Tracking and Occupational Attainment in Italy, in *Longitudinal and Life Course Studies*, 12(3):441-462.

Becker, W., Rosen, S. (1992): The Learning Effect of Assessment and Evaluation in High School, in *Economics of Education Review*, 11(2):107–118.

Betts, J.R. (1997): Do Grading Standards Affect the Incentive to Learn?, University of California at San Diego, Department of Economics, Discussion Paper 97-22.

Betts, J.R. (1998): The impact of educational standards on the level and distribution of earnings, in *American Economic Review*, 88:266-275.

Betts, J.R., Grogger, J. (2003): The impact of grading standards on student achievement, educational attainment, and entry-level earnings, in *Economics of Education Review*, 22: 343-352.

Birkelund, J.F., van de Werfhorst, H.G. (2022): Long-term Labor Market Returns to Upper Secondary School Track Choice: Leveraging Idiosyncratic Variation in Peers' Choices, in *Social Science Research*, 102:102629.

Bonesrønning, H. (1999): The Variation in Teachers' Grading Practices: Causes and Consequences, in *Economics of Education Review*, 18:89–105.

Bonesrønning, H. (2004): Do the Teachers' Grading Practices Affect Student Achievement?, in *Education Economics*, 12(2):151-167.

Bonner, S.M., Chen P.P. (2019): Chapter 3: The Composition of Grades: Cognitive and Noncognitive Factors, in Guskey T.R., Brookhart S.M. (Eds): *What We Know*

about Grading: What Works, What Doesn't, and What's Next, ASCD, Alexandia (VA).

Borghans L., Golsteyn B.H.H., Heckman J.J., Humphries J.E. (2016): What Grades and Achievement Tests Measure, in *PNAS*, 113(47):13354–13359.

Bracci, E. (2009): Autonomy, Responsibility and Accountability in the Italian School System, in *Critical Perspective on Accounting*, 20(3):293–312.

Cinelli, C., Forney, A., Pearl, J. (2021): A Crash Course in Good and Bad Controls, in *Sociological Methods & Research*, online.

Cherry, T.L., Ellis, L.V. (2005): Does Rank-Order Grading Improve Student Performance? Evidence From a Classroom Experiment, in *International Review of Economic Education*, 4(1):9–19.

Chetty R., Friedman J. N., Rockoff J. E. (2014): Measuring the Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood, in *American Economic Review*, 104(9):2633–79.

Chowdhury, F. (2018): Grade Inflation: Causes, Consequences and Cure, in *Journal of Education and Learning*, 7(6).

Chulkov, D.V. (2006): Student Response to Grading Incentives: Evidence from College Economics Courses, in *Journal of Instructional Psychology*, 33(3):206-211.

Clark, D.C. (1969): Competition for Grades and Graduate Student Performance, in *The Journal of Educational Research*, 62:351–354.

Correa, H., Gruver, G.W. (1987): Teacher-Student Interaction: A Game Theoretic Extension of the Economic Theory of Education, in *Mathematical Social Science*, 13:19-47.

Eccles, J.S., Wong, C.A., Peck, S.C. (2006): Ethnicity as a Social Context for the Development of African-American Adolescents, in *Journal of School Psychology*, 44(5):407–426.

Elwert, F., Winship, C. (2014): Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable, in *Annual Review of Sociology*, 40(1):31-53.

Facchinello, L. (2020): Short- and Long-run Effects of Early Grades, Available at SSRN: https://ssrn.com/abstract=2966571.

Fallan, L., Opstad, L. (2012): Attitude towards Study Effort Response to Higher Grading Standards: Do Gender and Personality Distinctions Matter?, in *Journal of Education and Learning*, 1(2):179-187.

Figlio, D.N., Lucas, M.E. (2004): Do High Grading Standards Affect Student Performance?, in *Journal of Public Economics*, 88:1815-1834.

Finefter-Rosenbluh, I., Levinson, M. (2015): What is Wrong with Grade Inflation (if Anything)?, in *Philosophical Inquiry in Education*, 23(1):3-21.

Gershenson, S. (2016): Linking Teacher Quality, Student Attendance, and Student Achievement, in *Education Finance and Policy*, 11(2):125–49.

Gillen-O' Neel, C., Ruble, D., Fuligni, A. (2011): Ethnic Stigma, Academic Anxiety, and Intrinsic Motivation in Middle Childhood, in *Child Development*, 82(5):1470–1485.

Gold, R.M., Reilly, A., Silberman, R., Lehr, R. (1971): Academic Achievement Declines Under Pass-Fail Grading, in *Journal of Experimental Education*, 39(3):17–21.

Hales, I.W., Bain, P.T, Rand, L.P. (1971): *An Investigation of Some Aspects of the Pass-Fail Grading System*, Annual Meeting of the American Educational Research Association, New York.

Hattie, J. (2008): *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*, Routledge.

Heckman, J.J., Kautz, T. (2014): Fostering and Measuring Skills: Interventions that Improve Character and Cognition, in Heckman, J.J., Humphries, J.E., Kautz, T. (Eds): *The Myth of Achievement Tests: The GED and the Role of Character in American Life*, Univ of Chicago Press, Chicago, 341–430.

Hernán, M.A. Robins, J.M. (2006): Instruments for Causal Inference: An Epidemiologist's Dream?, *Epidemiology*, 17(4):360-372.

Iacus, S.M., Porro, G. (2008): Teachers' Evaluations and Students' Achievement: How to Identify Grading Standards and Measure their Effects, in *Education Economics*, 19(2):139-159.

Jalava, N., Joensen, J.S., Pellas, E. (2015): Grades and Rank: Impacts of Non-financial Incentives on Test Performance, in *Journal of Economic Behavior & Organization*, 115:161-196.

Levitt, S.D., List, J.A., Neckermann, S., Sadoff, S. (2012): The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance, in NBER Working Paper (18165).

Manganelli, S., Cavicchiolo, E., Lucidi, F., Galli, F., Cozzolino, M., Chirico, A., Alivernini, F. (2021): Differences and Similarities in Adolescents' Academic Motivation Across Socioeconomic and Immigrant Backgrounds, in *Personality and Individual Differences*, 182:111077.

Montmarquette, C., Mahseredjian, S. (1989): Could Teacher Grading Practices Account for the Unexplained Variation in School Achievements?, in *Economics of Education Review*, 8(4):335–343.

Pei, Z., Pischke, J., Schwandt, H. (2019): Poorly Measured Confounders are More Useful on the Left than on the Right, in *Journal of Business & Economic Statistics*, 37(2):205-216.

Seaman, S.R., White, I.R. (2013): Review of Inverse Probability Weighting for Dealing with Missing Data, in *Statistical Methods in Medical Research*, 22(3):278–295.

Triventi, M. (2020): Are Children of Immigrants Graded Less Generously by their Teachers than Natives, and Why? Evidence from Student Population Data in Italy, in *International Migration Review* 54(3):765–795.

Triventi, M., Barone, C., Facchini, M. (2021): Upper Secondary Tracks and Student Competencies: A Selection or a Causal Effect? Evidence from the Italian Case, in *Research in Social Stratification and Mobility*, 76:100626.

Tyner A., Gershenson S. (2020): Conceptualizing Grade Inflation, in *Economics of Education Review*, 78.

Urhahne, D., Wijnia, L. (2021): A Review on the Accuracy of Teacher Judgments, in *Educational Research Review,* 32:100374.

Vecchione, M., Alessandri, G., Marsicano, G. (2014): Academic Motivation Predicts Educational Attainment: Does Gender Make a Difference?, in *Learning and Individual Differences,* 32:124-131.

Walvoord, B., Anderson, V.J. (1998): *Effective Grading: A Tool for Learning and Assessment,* San Francisco, CA: Jossey-Bass.

Appendix

Table A1: Description of the variables of interest

Control variables	Coding and description
Student sociodemographic & performance	
Gender	Recoded as 0 = Male; 1 = Female
Immigrant status	Recoded as 0 = Native; 1 = Immigrant I and II generation
ESCS	Standardized index from INVALSI composed by: parental occupation status, parental level of education, possession of specific material assets
Quarter of birth	Recoded as 0 = 1st quarter; 1 = 2nd quarter; 2 = 3rd quarter; 3 = 4th quarter
Regularity in studies	Recoded as 0 = Regular/early entrance; 1 = Late entrance
Attendance to infant school	Recoded as 0 = Yes; 1 = No; 2 = Missing
Attendance to kindergarten	Recoded as 0 = Yes; 1 = No; 2 = Missing
Student previous performance in (subject)	Grade at the end of 4th grade, self-reported (scale from 0 = 5 or less, to 5 = 10)
Teacher characteristics	
Gender	Recoded as 0 = Male; 1 = Female
Age	Continuous variable (scale from 25 to 68 in Mathematic; scale from 26 to 68 in Language)
Within-school seniority	Continuous variable (scale from 0 to 42 for Mathematics; scale from 0 to 41 for Language)
Educational credentials	Recoded as 0 = Teaching diploma; 1 = Bachelor/master degree/PhD
Parental education	Recoded as 0 = Lower; 1 = Higher
Type of contract	Recoded as 0 = Fixed-term; 1= Permanent
Teaching to test INVALSI (homework)	Recoded as 0 = No; 1 = Yes
Teaching to test INVALSI (in class)	Recoded as 0 = No; 1 = Yes
Classroom composition	
Share of female students	Continuous variable (scale from 0 to 100)
Mean ESCS (net of individual)	Standardized continuous variable
Classroom size	Continuous variable (scale from 11 to 29)
Share of immigrant students	Continuous variable (scale from 0 to 100)

Table A2: Descriptive statistics of the variables used in the analysis (N= 9,370)

Variable	**Mean**	**Std. Dev.**	**Min/Max**
Grading Standards in Language	0	1	-2.585/4.112
Grading Standards in Mathematics	0	1	-2.439/4.152
Language competence (test scores) in grade 8	0	1	-3.935/4.274
Language competence (test scores) in grade 10	0	1	-4.288/3.109
Mathematics competence (test scores) in grade 8	0	1	-4.364/4.098
Mathematics competence (test scores) in grade 10	0	1	-3.415/2.726
School track enrolment	0.381 0 = Non-academic track (61.94%) 1 = Traditional lyceum (38.06%)	0.486	0/1
Students' sociodemographic & performance			
Gender	0.52 0 = Male (47.99%) 1 = Female (52.01%)	0.5	0/1
Immigrant status	0.064 0 = Native (93.59%) 1 = 1st or 2nd gen. immigrant (6.41%)	0.245	0/1
ESCS	0.164	0.952	-2.701/2.429
Quarter of birth	1.514 0 = 1st quarter (23.5%) 1 = 2nd quarter (25.81%) 2 = 3rd quarter (26.51%) 3 = 4th quarter (24.18%)	1.097	0/3
Regularity in studies	0.013 0 = Regular/early entrance (98.74%) 1 = Late entrance (1.26%)	0.112	0/1
Attendance to infant school	0.946 0 = Yes (25.58%) 1 = No (54.28%) 2 = Missing (20.14%)	0.674	0/2
Attendance to kindergarten	0.131 0 = Yes (87.09%) 1 = No (12.69%) 2 = Missing (0.22%)	0.344	0/2
Previous performance in Language Previous performance in Mathematics	3.611 3.673	1.026 1.056	0/5 0/5
Language teacher characteristics			
Gender	0.978 0 = Male (2.17%) 1 = Female (97.83%)	0.146	0/1
Age	51.894	7.722	26/68

Within-school seniority	15.48	8.890	0/41
Educational credentials	0.27 0 = Teaching diploma (72.97%) 1 = Bachelor/Master/PhD (27.03%)	0.444	0/1
Parental education	0.356 0 = Lower (64.41%) 1 = Higher (35.59%)	0.479	0/1
Type of contract	0.952 0 = Fixed-term (4.83%) 1 = Permanent (95.17%)	0.215	0/1
Teaching to test (homework)	0.718 0 = No (28.15%) 1 = Yes (71.85%)	0.450	0/1
Teaching to test (in class)	0.164 0 = No (83.56%) 1 = Yes (16.44%)	0.371	0/1
Mathematics teacher characteristics			
Gender	0.974 0 = Male (2.57%) 1 = Female (97.43%)	0.158	0/1
Age	51.571	8.121	25/68
Within-school seniority	14.670	8.720	0/42
Educational credentials	0.275 0 = Teaching diploma (72.48%) 1 = Bachelor/Master/PhD (27.52%)	0.447	0/1
Parental education	0.361 0 = Lower (63.85%) 1 = Higher (36.15%)	0.480	0/1
Type of contract	0.947 0 = Fixed-term (5.33%) 1 = Permanent (94.67%)	0.225	0/1
Teaching to test (homework)	0.704 0 = No (29.55%) 1 = Yes (70.45%)	0.456	0/1
Teaching to test (in class)	0.172 0 = No (82.85%) 1 = Yes (17.15%)	0.377	0/1
Classroom composition			
Share of female students	49.877	11.322	8.333/92.857
Mean ESCS (net of individual)	0.090	0.533	-1.477/2.098
Classroom size	18.293	3.647	11/29
Share of immigrant students	8.922	12.687	0/100

Table A3: Balancing Tests: as-good-as-random distribution of Grading Standards across students in Language

	Gender		***Ethnic status***		***Socio-economic origin***	
Grading Standard in *Language*	-0.01	0	0	0	0.07**	0.07**
	(0.01)	(0.02)	(0.00)	(0.01)	(0.01)	(0.03)
Constant	0.52***	0.52***	0.06***	0.06***	0.16***	0.16***
	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)	(0.01)
R-sqr	0	0.05	0	0.19	0	0.26
F-Statistic	0.29	0.89	0.64	0.63	0	0.01
BIC	13603.42	13104.42	251.51	-1711.51	25645.19	22844.7
AIC	13589.12	13090.13	237.22	-1725.8	25630.9	22830.41
Obs.	9370	9370	9370	9370	9370	9370
School FE	NO	YES	NO	YES	NO	YES

Table A4: Balancing Tests: as-good-as-random distribution of Grading Standards across students in Mathematics

	Gender		***Ethnic Status***		***Socio-economic origin***	
Grading Standard in *Mathematics*	-0.01	-0.01	0	0	0.04**	0.06**
	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.02)
Constant	0.52***	0.52***	0.06***	0.06***	0.16***	0.16***
	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)	(0.01)
R-sqr	0	0.05	0	0.19	0	0.26
F-Statistic	0.08	0.42	0.73	0.58	0	0.01
BIC	13601.39	13103.76	251.61	-1711.58	25672.02	22844.67
AIC	13587.1	13089.47	237.32	-1725.87	25657.73	22830.38
Obs.	9370	9370	9370	9370	9370	9370
School FE	NO	YES	NO	YES	NO	YES

Table A5: Comparison between linear and quadratic regressions predicting INVALSI test score in Language and Mathematics in 8th grade. Coefficients derived from model 3 (all controls + fixed effects at the school level). Standard error in parentheses; ***p<0.01, **p<0.05, *p<0.1

	Language 8th grade	(S.E.)	Language 8th grade	(S.E.)	Mathematics 8th grade	(S.E.)	Mathematics 8th grade	(S.E.)
Grading Standards (5th grade)	0.133***	(0.029)	0.145***	(0.030)	0.090***	(0.027)	0.108***	(0.029)
Grading Standards ^2			-0.026	(0.016)			**-0.028***	(0.016)
Student Characteristics								
Female (Ref. Male)	0.263***	(0.020)	0.263***	(0.020)	-0.183***	(0.021)	-0.181***	(0.021)
Quarter of birth (Ref. 1st)								
2nd quarter	0.073***	(0.028)	0.074***	(0.028)	0.067**	(0.028)	0.067**	(0.028)
3rd quarter	-0.006	(0.028)	-0.005	(0.028)	-0.018	(0.028)	-0.017	(0.028)
4th quarter	0.004	(0.028)	0.004	(0.028)	0.000	(0.029)	-0.000	(0.029)
ESCS	0.215***	(0.012)	0.216***	(0.012)	0.168***	(0.012)	0.167***	(0.012)
Immigrant (Ref. Native)	-0.182***	(0.045)	-0.180***	(0.045)	-0.106**	(0.046)	-0.107**	(0.046)
Late entrance (Ref. Regular)	-0.122	(0.091)	-0.122	(0.091)	-0.060	(0.093)	-0.060	(0.093)
Attendance to infant school (Ref. Yes)								
No	-0.005	(0.025)	-0.005	(0.025)	0.003	(0.026)	0.003	(0.026)
Missing	-0.028	(0.050)	-0.027	(0.050)	-0.044	(0.051)	-0.044	(0.051)
Attendance to kindergarten (Ref. Yes)								
No	-0.123*	(0.067)	-0.120*	(0.067)	-0.170**	(0.068)	-0.170**	(0.068)
Missing	-0.738***	(0.209)	-0.737***	(0.209)	-0.597***	(0.212)	-0.600***	(0.212)
Teacher Characteristics								
Female (Ref. Male)	-0.134	(0.118)	-0.136	(0.118)	-0.300**	(0.130)	-0.260**	(0.132)
Age	0.001	(0.003)	0.002	(0.003)	0.002	(0.003)	0.003	(0.003)
Seniority in school (years)	-0.005**	(0.003)	-0.006**	(0.003)	-0.004	(0.003)	-0.004	(0.003)
Bachelor/Master/PhD (Ref. Teaching diploma)	0.003	(0.044)	0.005	(0.044)	-0.122***	(0.044)	-0.127***	(0.044)
Parental education higher (Ref. Lower)	-0.005	(0.039)	-0.005	(0.039)	0.012	(0.038)	0.013	(0.038)
Permanent contract (Ref. Fixed-term)	0.311***	(0.092)	0.309***	(0.092)	0.033	(0.091)	0.038	(0.091)
Teaching to test in class yes (Ref. No)	-0.032	(0.045)	-0.031	(0.045)	-0.058	(0.044)	-0.056	(0.044)
Teaching to test homework yes (Ref. No)	-0.004	(0.053)	-0.010	(0.053)	0.028	(0.049)	0.024	(0.049)
Classroom Composition								
% Female	-0.001	(0.001)	-0.001	(0.001)	0.001	(0.001)	0.002	(0.002)
Mean ESCS	0.023	(0.053)	0.027	(0.053)	-0.045	(0.054)	-0.057	(0.054)
Class size	-0.002	(0.006)	-0.002	(0.006)	-0.002	(0.006)	-0.002	(0.006)
% Immigrants	0.001	(0.002)	0.001	(0.002)	-0.001	(0.002)	-0.002	(0.002)
Constant	-0.074	(0.247)	-0.088	(0.247)	0.456*	(0.242)	0.413*	(0.243)
Observations	9,370		9,370		9,370		9,370	
R-squared	0.191		0.191		0.202		0.202	
AIC	24351.71		24350.88		24658.88		24657.37	
BIC	24530.34		24536.65		24837.51		24843.14	

Table A6: Comparison between linear and quadratic regressions predicting INVALSI test score in Language and Mathematics in 10th grade. Coefficients derived from model 3 (all controls + fixed effects at the school level). Standard error in parentheses; ***p<0.01, **p<0.05, *p<0.1

	Language		Language		Mathematics		Mathematics	
	10th grade	(S.E.)	10th grade	(S.E.)	10th grade	(S.E.)	10th grade	(S.E.)
Grading Standards (5th grade)	0.114***	(0.029)	0.136***	(0.030)	0.097***	(0.026)	0.105***	(0.028)
Grading Standards ^2			**-0.052*** **	(0.016)			-0.013	(0.015)
Student Characteristics								
Female (Ref. Male)	0.251***	(0.020)	0.251***	(0.020)	-0.219***	(0.020)	-0.218***	(0.020)
Quarter of birth (Ref. 1st)								
2nd quarter	0.045*	(0.027)	0.047*	(0.027)	0.021	(0.027)	0.021	(0.027)
3rd quarter	-0.033	(0.027)	-0.032	(0.027)	-0.039	(0.027)	-0.039	(0.027)
4th quarter	0.020	(0.028)	0.021	(0.028)	-0.021	(0.028)	-0.021	(0.028)
ESCS	0.244***	(0.012)	0.245***	(0.012)	0.216***	(0.011)	0.216***	(0.011)
Immigrant (Ref. Native)	-0.142***	(0.045)	-0.137***	(0.045)	-0.077*	(0.044)	-0.078*	(0.044)
Late entrance (Ref. Regular)	-0.133	(0.091)	-0.133	(0.091)	-0.050	(0.090)	-0.049	(0.090)
Attendance to infant school (Ref. Yes)								
No	-0.009	(0.025)	-0.009	(0.025)	-0.010	(0.025)	-0.010	(0.025)
Missing	0.069	(0.050)	0.071	(0.050)	0.051	(0.049)	0.051	(0.049)
Attendance to kindergarten (Ref. Yes)								
No	-0.122*	(0.066)	-0.116*	(0.066)	-0.154**	(0.066)	-0.154**	(0.066)
Missing	-0.831***	(0.207)	-0.830***	(0.207)	-0.921***	(0.205)	-0.922***	(0.205)
Teacher Characteristics								
Female (Ref. Male)	-0.162	(0.117)	-0.167	(0.117)	-0.047	(0.125)	-0.028	(0.127)
Age	-0.002	(0.003)	0.000	(0.003)	-0.000	(0.003)	-0.000	(0.003)
Seniority in school (years)	0.000	(0.003)	0.000	(0.003)	0.000	(0.003)	-0.000	(0.003)
Bachelor/Master/PhD (Ref. Teaching diploma)	-0.037	(0.044)	-0.033	(0.044)	-0.086**	(0.043)	-0.088**	(0.043)
Parental education higher (Ref. Lower)	-0.006	(0.038)	-0.006	(0.038)	0.028	(0.037)	0.029	(0.037)
Permanent contract (Ref. Fixed-term)	0.043	(0.091)	0.040	(0.091)	0.014	(0.087)	0.017	(0.087)
Teaching to test in class yes (Ref. No)	0.038	(0.044)	0.039	(0.044)	-0.026	(0.043)	-0.025	(0.043)
Teaching to test homework yes (Ref. No)	-0.020	(0.052)	-0.031	(0.052)	-0.007	(0.048)	-0.009	(0.048)
Classroom Composition								
% Female	0.001	(0.001)	0.001	(0.001)	-0.002	(0.233)	-0.022	(0.235)
Mean ESCS	0.042	(0.052)	0.050	(0.052)	-0.002	(0.052)	-0.008	(0.052)
Class size	0.011**	(0.005)	0.010*	(0.005)	0.009*	(0.005)	0.009*	(0.005)
% Immigrants	-0.003	(0.002)	-0.002	(0.002)	-0.001	(0.002)	-0.001	(0.002)
Constant	-0.182	(0.245)	-0.209	(0.245)	-0.002	(0.233)	-0.022	(0.235)
Observations	9,370		9,370		9,370		9,370	
R-squared	0.197		0.198		0.231		0.231	
AIC	24172.41		24163.16		23971		23972.17	
BIC	24351.04		24348.93		24149.63		24157.95	

Table A7: Comparison between linear and quadratic regressions predicting students' enrollment in traditional lyceums in 10th grade for Language grading standards and Mathematics grading standards. Coefficients derived from model 3 (all controls + fixed effects at the school level). Standard error in parentheses; ***p<0.01, **p<0.05, *p<0.1

	Trad. Lyceum (Lang)	(S.E.)	Trad. Lyceum (Lang.)	(S.E.)	Trad. Lyceum (Maths)	(S.E.)	Trad. Lyceum (Maths)	(S.E.)
Grading standards (5th grade)	0.027*	(0.014)	0.031**	(0.015)	0.041***	(0.013)	0.049***	(0.014)
Grading standards ^2			-0.009	(0.008)			-0.011	(0.007)
Student Characteristics								
Female (Ref. Male)	-0.026***	(0.010)	-0.026***	(0.010)	-0.026***	(0.010)	-0.025**	(0.010)
Quarter of birth (Ref. 1st)								
2nd quarter	0.024*	(0.013)	0.025*	(0.013)	0.024*	(0.013)	0.024*	(0.013)
3rd quarter	-0.000	(0.013)	-0.000	(0.013)	-0.001	(0.013)	-0.001	(0.013)
4th quarter	-0.007	(0.014)	-0.007	(0.014)	-0.007	(0.014)	-0.008	(0.014)
ESCS	0.138***	(0.006)	0.139***	(0.006)	0.139***	(0.006)	0.139***	(0.006)
Immigrant (Ref. Native)	-0.054**	(0.022)	-0.053**	(0.022)	-0.051**	(0.022)	-0.052**	(0.022)
Late entrance (Ref. Regular)	-0.044	(0.045)	-0.044	(0.045)	-0.045	(0.045)	-0.045	(0.045)
Attendance to infant school (Ref. Yes)								
No	-0.018	(0.012)	-0.018	(0.012)	-0.017	(0.012)	-0.017	(0.012)
Missing	-0.001	(0.025)	-0.001	(0.025)	-0.001	(0.025)	-0.001	(0.025)
Attendance to kindergarten (Ref. Yes)								
No	0.045	(0.033)	0.046	(0.033)	0.039	(0.033)	0.040	(0.033)
Missing	-0.197*	(0.102)	-0.197*	(0.102)	-0.197*	(0.102)	-0.198*	(0.102)
Teacher Characteristics								
Female (Ref. Male)	-0.135**	(0.058)	-0.136**	(0.058)	-0.058	(0.062)	-0.042	(0.063)
Age	0.000	(0.001)	0.001	(0.001)	0.002	(0.001)	0.002	(0.001)
Seniority in school (years)	0.000	(0.001)	0.000	(0.001)	-0.002	(0.001)	-0.002*	(0.001)
Bachelor/Master/PhD (Ref. Teaching diploma)	0.003	(0.021)	0.004	(0.021)	0.030	(0.021)	0.028	(0.021)
Parental education higher (Ref. Lower)	0.016	(0.019)	0.016	(0.019)	0.048***	(0.018)	0.048***	(0.018)
Permanent contract (Ref. Fixed-term)	0.068	(0.045)	0.067	(0.045)	-0.013	(0.043)	-0.011	(0.043)
Teaching to test in class yes (Ref. No)	-0.010	(0.022)	-0.010	(0.022)	-0.033	(0.021)	-0.032	(0.021)
Teaching to test homework yes (Ref. No)	-0.040	(0.026)	-0.042	(0.026)	-0.007	(0.024)	-0.009	(0.024)
Classroom Composition								
% Female	-0.000	(0.001)	-0.000	(0.001)	-0.000	(0.001)	-0.000	(0.001)
Mean ESCS	0.091***	(0.026)	0.092***	(0.026)	0.093***	(0.026)	0.088***	(0.026)
Class size	0.004	(0.003)	0.004	(0.003)	0.005*	(0.003)	0.005*	(0.003)
% Immigrants	0.001	(0.001)	0.001	(0.001)	0.001	(0.001)	0.001	(0.001)
Constant	0.360***	(0.120)	0.355***	(0.121)	0.273**	(0.116)	0.256**	(0.117)
Observations	9,370		9,370		9,370		9,370	
R-squared	0.211		0.212		0.212		0.212	
AIC	10875.12		10875.6		10866.46		10866.01	
BIC	11053.75		11061.38		11045.1		11051.79	

Figure A1: Plot of predicted values derived from quadratic regression models predicting INVALSI test score in Mathematics in 8th grade. Coefficients derived from model 3 (all controls + fixed effects at the school level).

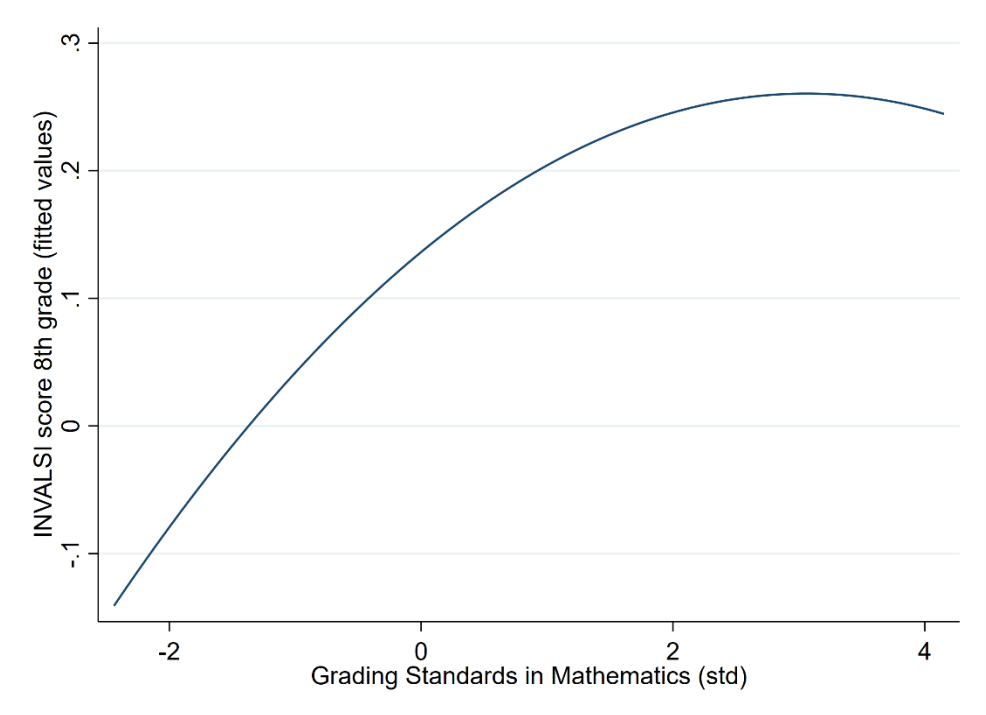

Figure A2: Plot of predicted values derived from quadratic regression models predicting INVALSI test score in Language in 10th grade. Coefficients derived from model 3 (all controls + fixed effects at the school level).