Title: The Impact of Discrimination on Future Performance

Subtitle: evidence of a self-fulfilling prophecy from the Italian student population

Author : Paolo Agnolin

Thesis Director : William Parienté

Thesis Reader : Tommaso Frattini

Academic Year 2019-2020

In order to obtain the joint degree:

Master 120 en Sciences économiques, Orientation générale, Finalité approfondie (UCL/UNamur)

and

Dottore magistrale in Economics and Political Science (UNIMI)

Université catholique de Louvain - Economics School of Louvain - Place Montesquieu 3, 1348 Louvain-La-Neuve
Université de Namur – Economics School of Namur – Rempart de la Vierge 8, 5000 Namur
Università degli Studi di Milano • Via Festa del perdono 7 • 20122 Milano / Italy

THE IMPACT OF DISCRIMINATION ON FUTURE PERFORMANCE:

evidence of a self-fulfilling prophecy from the Italian student population*

Paolo Agnolin†

2020
August

Abstract

If foreign students are affected by teacher's biased assessment, is there an impact on their long-term future school competence? We study this question using educational data on the entire population of Italian students and we find evidence of grading bias against students with a foreign status. In fact, teachers give lower grades to foreign pupils compared to natives who have the same performance on standardized, blindly-graded, tests. We then estimate the impact of grading under-assessment over the future performance of affected students, and we find significant negative effects. Grading bias seems to lead to a self-fulfilling prophecy by influencing the behaviour of discriminated groups in the direction of teacher assessment.

*I would like to thank prof. William Parienté (Université Catholique de Louvain) for his competent and inspiring supervision. I thank INVALSI for the data and especially Patrizia Giannantoni (INVALSI) for the fundamental support. I am also deeply thankful to Adam Levai (Université Catholique de Louvain) and Alain Guillet (Université Catholique de Louvain) for strategic assistance in data management and storage, as well as to prof. Michela Carlana (Harvard University), prof. Ernesto San Martín Gutiérrez (Pontifical Catholic University of Chile) and prof. Moris Triventi (University of Trento) for their useful comments and suggestions. Finally, a special thank goes to prof. Fabrizio Iacone (University of Milan), prof. Tommaso Frattini (University of Milan), prof. Sébastian Van Bellegem (Université Catholique de Louvain) and again prof. William Parienté (Université Catholique de Louvain) for the inspirational econometric training.

†Economics School of Louvain, Université Catholique de Louvain and Department of Economics, Management and Quantitative Methods, University of Milan.

1 Introduction

Equality of educational opportunity is not just a widely accepted indicator of social mobility in societies (Grusky and Hauser, 1984), but also one of the main goals toward which any education system should tend. A condition of equality in the school system is often seen as a prerequisite of social cohesion, economic performance and quality of democracy. Educational inequality is usually defined as the dependence of people educational achievement on their social background. A large body of social science literature has established the economic, educational, and migration status as robust and highly influential determinants of children's educational attainments (Sewell and Shah, 1967; Schlicht-Schmälzle and Ackermann, 2012).

Although socio-economic background and parental education are not of less importance, the migration status of the student becomes more and more relevant for equality of opportunity in western societies with increasing shares of international migration. This is especially true in Italy, where mass immigration is a relatively recent phenomenon and where the growing number of immigrant students has profoundly and rapidly changed the challenges that the national schooling system must face in order to ensure skill development in a diverse student population and promote social cohesion. While in western countries the negative impact of a migration status on educational achievement is an uncontroversial fact in demographic and education literature, after controlling for parental socioeconomic and educational status the causal link between migration background and educational success is still unsettled.

In this study, we[1] investigate one of the possible specific mechanisms through which inequality of educational opportunity could operate and propagate. This is the presence of implicit grading bias that may lead teachers to systematically under-assess the competence of students with a migration background. Stereotypes, which are over-generalized representations of characteristics of certain groups (Bordalo et al., 2016), are not always an explicit phenomenon and allow for easier and efficient processing of information. On the other hand, they may cause biased judgment or even discrimination against some groups, such as students with different ethnicity. This process is not neutral to the future propagation of inequality of educational achievements, as discrimination may lead to self-fulfilling prophecies by influencing the behavior of discriminated groups in the direction of the stereotypes (Glover et al., 2017).

Our empirical analysis focuses on the early stages of the scholastic experience by using educational data collected on the entire Italian population of students enrolled in the second grade (grade II) of primary education in the academic year 2012/2013. We carry out a multi-step merging process of different cross-sections provided by the National Institute for the Evaluation of the Italian School System (INVALSI) that allows us to track individuals

[1]The first plural person is used here in the farthest possible way to the plurale maiestatis. In fact, it is adopted in order to acknowledge the fundamental contribution that other people have brought to this research, in primis prof. William Parienté (Université Catholique de Louvain). Although this research has benefited from key comments from other scholars, all the possible mistakes are mine.

over three periods from 2012 to 2018.

We begin by estimating a measure of the intensity of the grading bias experienced by the students and of the average teacher bias in the class. We adopt an empirical approach that has become common in the recent literature about grading bias, which is the comparison of student assessment in blind and non-blind tests. We see that, holding constant performance on standardized, blindly-graded tests, immigrant students receive lower grades when graded by their teachers in a non-blind way compared to native peers in the class. Unfortunately, we are unable to completely rule out the hypothesis that part of our final grading bias measure could partially capture some other determinants of individual under-assessment in addition to those we control for. However, our results are consistent to the previous literature about grading bias, such as Alesina et al. (2018) and Triventi (2019), who also exploit information about an Implicit Association Test and about students' behaviour and commitment.

In fact, this is not the first study to estimate grading bias in the Italian school system, and the work of Alesina et al. (2018) and Triventi (2019) should be considered as important robust attempts to solve our identification problem. However, the contribution regarding the future effects of grading bias taking place at the early stages of education remains rather scarce. To our knowledge, this is the first work that exploits this measure of ethnic-based grading bias to estimate the impact on future school competence. This is a key research question to answer, as intra-classroom evaluation biases may very well lead to gaps in future scholastic attainments by ultimately realizing the unfavourable self-fulfilling prophecy of discrimination.

We find that keeping fixed competence and other characteristics in the first observed period, students who were affected by grading bias eventually perform worse in the next periods compared to similar peers that did not experience intra-classroom bias. Minority students who experience under-assessment or that are taught by biased teachers tend to be characterised more likely by an irregular school path and to present lower performance in future national and blindly-graded tests. Our main results are robust to different model specifications.

The reminder of this work is organized as follows. Section 2 summarizes the relevant previous literature about the matter in hand. Section 3 provides some background information about the Italian context and foreign students in the national school system. Section 4 describes our data and how the grading bias measures are constructed. Section 5 presents our main results and section 6 concludes.

2 Literature review

The literature regarding discrimination in school and grading bias has been flourishing in the last decades and many relevant publications have contributed to shed light on these topics.

In general, there are two main reasons for teachers to systematically misevaluate the performance of students with certain characteristics. The first is that teachers may merely like or dislike people with some traits, and accordingly imposing rewards or punishments ("taste discrimination"). Instead, the second implies that teachers may be characterised by a more sophisticated bias, evaluating student performance by also using observed characteristics that, according to their mental categories and their experience, are perceived to be correlated with competence ("statistical discrimination"). The literature that this research is aimed to expand usually refer to the latter concept.

There is a large body of recent evidence showing that statistical discrimination in grading is at work in both developing and developed countries, although there are also contrasting results. Van Ewiik (2010) finds no evidence of discrimination in primary education in the Netherlands, while other research reports signs of discrimination against students with foreign surnames in Germany (Sprietsma, 2013) and Russia (Akifyeva and Alieva, 2016). Although Van Ewijk (2010) finds no evidence that teachers grade minority and majority students differently for the same performance, the study highlights teachers' lower expectations and unfavourable attitudes that both likely affect their behaviour towards minority students, potentially inducing them to under-perform. The effect over minority grades thus seems to be indirect. Another relevant field experiment is run in India by Hanna and Linden (2009), who randomly assign identifying cover pages to exam scripts and compare the marks with and without these cover sheets. They find evidence of significant discrimination with exams assigned to lower caste children being given lower grades than those assigned to higher caste children. Gilliam et al. (2016) run a laboratory experiment, where they track teachers' eye gazes while watching a video and find that, when expecting challenging behaviours, teachers gazed longer at black children, even if all children were behaving similarly.

Most research trying to infer teacher discrimination from observational data finds evidence of discrimination against pupils with a migration background (Alesina et al., 2018; Botelho et al., 2015; Carlana et al., 2018; Hinnerich et al. 2015; Triventi, 2019), while others find none or even a premium for specific categories (Lindahl 2007; Burgess and Greaves 2013). This kind of studies often exploits the comparison between teacher-assigned grades and standardized-blind test scores in order to infer the presence of evaluation bias (Botelho et al., 2015; Burgess and Greaves, 2013; Hanna and Linden, 2012; Lavy, 2008; Lavy and Megalokonomou, 2019 ; Lavy and Sand, 2018; Van Ewijk, 2011). This is the approach that will be adopted also in this research.

Botelho et al. (2015) use observational data from Brazil and, after holding constant performance in blindly-scored tests of proficiency and behavioural traits, they find that teacher-assigned grades suffer from cardinal and ordinal biases. Similarly to Altonji and Pierret (2001), they also examine whether the duration of interaction between teachers and students produces different patterns, finding positive results. The basic idea is that the longer pupils and teachers interact, the smaller is the role of biased priors that emphasize racial identity.

Burgess and Greaves (2013) investigate differences in teacher grading according to ethnic background using observational data from England, finding significant under-assessment of Black Caribbean and Black African pupils, but also over-assessment of Asian students. Statistical discrimination comes from the process whereby teachers categorise students and create stereotypes or exemplars to make conscious or unconscious judgements about future students of the same group. Alesina et al. (2018) study the behaviour of teachers toward immigrant and native students in Italian middle schools and find evidence of teachers' bias in grading immigrants. The authors are able to distinguish between the role of teachers' biases and unobserved student characteristics, using an Implicit Association Test as a direct measure of teachers' stereotypes.

The study of the initial period of mandatory education is particularly important, since it is the time when parents invariably invest relying on the "asset-return" evaluations of more informed experts, i.e. teachers. *"For our purposes, the key element of this reasoning is that teacher communications may steer investment decisions in one way or the other. That is to say; parents (and children themselves) likely update investment (and effort) decisions after extracting information from report cards issued by teachers. Therefore, if children's perceived competence increases the returns or reduces the costs of investments, as in the traditional Beckerian human-capital framework, this mechanism can reinforce racial gaps in the accumulation of human capital. In this case, intra-classroom evaluation biases may very well lead to gaps in attainment, school choice, future scholastic performance and, ultimately, labor market outcomes"* (Botelho et al. 2015). This concern is confirmed by the study of Carlana et al. (2018), who analyse the educational choices of children of immigrants in a tracked school system. They show that immigrant boys in Italy enrol disproportionately into vocational high schools, as opposed to technical and academically oriented high schools, compared to natives of similar ability.

The interest of the introduced field of research is therefore not restricted to the identification of discrimination or grading bias evidence, but also to the effects that these processes generate on students' future behaviour, competence, and attainments. In fact, discrimination may lead to self-fulfilling prophecies by influencing the behaviour of discriminated groups in the direction of the stereotypes. Low expectations may lead to people reducing their effort at school, and therefore to achieving lower levels of human capital. For example, individuals exposed to bias toward their own group may reduce effort, self-confidence, and productivity (Glover et al., 2017). Statistical discrimination can potentially discourage minorities skill investments by implicitly persuading minorities that these investments would not be fully rewarded (Lundberg and Startz, 1983; Coate and Loury, 1993). The first to discuss this 'self-fulfilling prophecy' was probably Arrow (1973), while more recently it emerges again with Mechtenberg (2006) in the context of a cheap talk game between teachers and pupils. The propagation of negative stereotypes is part of the broad pattern of persistent inequalities, as argued by Hoff and Pandey (2006).

According to psychological literature, students from stigmatized groups and low-achieving students seem to be the more prone to this self-fulfilling prophecy, also called as "pygmalion effect" (Jussim and Harber, 2005; Rosenthal and Jacobson, 1965). Teachers holding negative attitudes or stereotypes toward an ethnic group may implicitly unveil their prejudices to the student through unintended changes in behavior. In order to observe this process at stake, stereotypes do not need to be explicit and teachers can be fully unaware of their bias. Casteel (1998) and Good (1987) provide two interesting examples of such unconscious processes. The former shows that white teachers call less on Afro American students than on white students to answer questions in class and help them less in finding the correct answer. The latter instead highlights that teachers are less demanding, provide less feedbacks, pay less attention and praise less often for success when they interact with students about whom they have lower expectations, the effect being weaker motivation and self-confidence in minority students which eventually translate into poorer performance. An alternative explanation of how minority students may be affected pertains what Dee (2005) calls "passive teacher effect". According to this framework, teacher behavior may remain unchanged but still students' behavior changes, in reaction to teacher demographic characteristics. A teacher with the same ethnicity of the students can become a role model for the students. On the other hand, if there is a difference between the ethnicity of the teacher and the pupil, and a stereotype about poorer performance of a minority exists, minority students will indeed have lower performances if they expect the teacher to share the negative stereotype.

While the psychological mechanism through which stereotypes affect students' confidence and future performance is still uncertain and it is a matter of psychological research, economic and statistical research can contribute to estimate the long-term impact of grading bias over some outcome variables, such as performance in standardized tests, learning results, school attainments, as well as future academic, professional and life choices. In this respect, Lavy and Sand (2018) and Lavy and Megalokonomou (2019) provide two pioneering contributions with respect to gender grading bias. The former study estimates the effect of primary school teachers' gender biases on boys' and girls' academic achievements during middle and high school and on the choice of advanced level courses in math and sciences during high school. Teachers' biased behavior at early stage of schooling actually seems to generate long-run implications for occupational choices and earnings at adulthood, with a heterogeneous impact, being larger for children from patriarchal families and for girls from low socioeconomic background (Lavy and Sand, 2018). Similarly, Lavy and Megalokonomou (2019) use data from a large number of school in Greece and find substantial effects of teacher biases on students' school attendance and performance in university admission exams, quality of enrolled degree and the given field of study at the university. This literature, however, investigates the effects of gender grading bias, while the focus of our research is on ethnic-based evaluation bias.

Although this is not the first study that estimates grading bias with respect to students with a migration background in Italy (see also Alesina et al., 2018, and Triventi, 2019), to our knowledge this the first research that estimates the future long-term impact of grading bias on migrants' school competence. This is a crucial issue with important policy implications, because the endogenous weaker educational progress would provide further evidence of how the propagation of negative stereotypes can crystallize a pattern of persistent inequalities.

3 Institutional background and the Italian context

3.1 Italian education system and INVALSI tests

In Italy, compulsory education starts at the age of 6, when pupils enter primary school. Before that moment, children can attend two preliminary cycles, that are the preschool (*asilo nido*) and the kindergarten (*scuola dell'infanzia*). The compulsory education system is composed of primary school (for children between the age of 6 and 11), lower secondary school (also called middle school, between the age of 11 and 14) and upper secondary school (between the age of 14 and 19). According to the Italian law that expects a minimum compulsory schooling, students can potentially conclude their study before the age of 19, provided that they attend at least 10 years of education (from age of 6 to 16) and that they achieve a qualification for upper secondary education that lasts at least 3 years. While the first two compulsory cycles (i.e. primary and lower secondary school) are characterized by an educational curriculum that is general, universal, and aimed at providing a common, basic, education to everyone, upper secondary education is characterised by different paths. In fact, students are free to choose between academic high school (*liceo*), technical high school and vocational high school. However, our research focuses on children in primary and middle school only, from grade II (i.e. the second grade in primary school) to grade VIII (i.e. the third and last year of middle school). During these years, students are assigned to the same class for all subjects, and on average classes are composed of around 20 students. Usually, teachers tend to be assigned to the same class for the whole cycle, although changes are possible, and teachers are always different between primary and lower secondary school. In primary school, students usually spend 7 or 8 hours with the Italian (language and literature) teacher and 6 hours with the Math teacher in grade I and II, while later they spend 6 hours in each of the two main subjects. In the middle school, they usually spend 6 hours in Italian and 5 in Math. Teachers are assigned to schools by the Italian Ministry of Education (MIUR) and their allocation is based on seniority and qualifications, with more experienced teachers that tend to be assigned to schools that are more preferred to them, as they tend to move closer to their home town and away from disadvantaged areas (Barbieri et al., 2011).

For assessment purposes, the school year is usually divided into four-month terms concluding in January and June. Students are assessed continuously, but official, periodic, and final evaluation of students' learning outcomes takes place two times a year, at the end of each term. This evaluation is expressed in numerical grades, that range from 1 to 10 (although grades below 4 are extremely rare), with 6 being the passing grade. Non-admission of students to the following school grade is relatively unlikely and must be unanimously decided by the class teaching staff. Failure can occur when the learning outcomes are largely unachieved, although teachers' decision is often affected by other criteria such as the number of absences and the presence of antagonistic conducts by the student.

While the teacher assessment is subjective and may also incorporate a broader evaluation of students' behavior, standardized and blindly-graded national tests take place every year in some specific grade. These are mainly multiple-choice exams in Italian, Math and English, administered by the National Institute for the Evaluation of the Italian Education System (INVALSI). These tests are blindly-graded on the base of a common and precise evaluation grid, and students' own teacher plays no role in the evaluation. INVALSI tests were administered to students in grade II and V (primary school), VI and VIII (corresponding to the first and third year of the lower secondary education), X and XIII (corresponding to the second and fifth year of the upper secondary education). However, since the academic year 2013/2014, INVALSI is no longer including grade VI, while the grade XIII is analysed only from the academic year 2018/2019, which means that in most of the years the grades where INVALSI evaluation takes place are only the second, fifth, eighth and tenth. The continuous student evaluation administered by INVALSI will be a key element in our analysis.

3.2 Immigrants in Italian schools

While since its unification Italy has always been a country of emigration[2], starting from the 1980s the immigration flows have also gained importance. Therefore, unlike the situation of other western countries, in Italy large-scale immigration is a relatively recent phenomenon. In 1981, the number of registered foreign residents was 321 000, in 1991 it was twice as large (around 625 000 people), and progressively scaled up to the 5,3 millions of 2020, which represent the 8,8% of the Italian population (ISTAT, 2020). Analogously, also the number of foreign children in the Italian school system has raised among the last decades, as shown in Figure 1.

Most of the foreign population comes from low and middle income countries such as Romania, Morocco, and Albania, and on average is characterized by lower socioeconomic status than the native population. On average, the foreign population is also younger and households tend to have more children. This translates into the fact that the proportion of foreign students over the total student population is higher than the whole percentage of foreign residents in the country (Carlana et al., 2018). Accordingly, in the academic year 2017/2018, the 9,7% of the student population was composed of foreign students. The

[2]From 1876 to 1976 around 24 millions of people left the country, with a maximum in 1913 (870 000 emigrants).

proportion is higher in primary school (11,2%), and it gradually decreases in the following school cycles[3] (MIUR, 2019), reflecting higher dropout rates of non-native students in later grades compared to natives.

The ethnic composition of immigrants' children in Italy is quite heterogeneous, as there is a substantial representation of various regions, such as East-Europe (especially Romania, Albania, Ukraine and Moldavia), North Africa (Morocco and Egypt), Asia (China, Philippines, India and Pakistan) and South America (Ecuador and Peru). However, more than half of foreign students come from only four countries, which are Romania, Albania, Morocco and China, with the first being the more represented (1/4 of the foreign student population). Figure 2 and Table 1 show the proportion of foreign students by nationality and their distribution in the different school levels.

Generally, in the Italian school system children of immigrants tend to be characterised by worst outcomes than their native peers, such as performance in standardized test scores[4] and grades at the end of the school cycles (INVALSI, 2019; Barban and White, 2012). Moreover, they lag behind native students in terms of academic performance and transitions to higher educational levels, with especially first-generation immigrant students that are frequently enrolled in classes at a lower grade than that of their age group, mainly because of language difficulties (Azzolini, 2011). The gap between native and foreign students also takes the form of more pronounced risks of delayed educational careers, lower university enrollment, and higher dropout risks, especially in high school (Azzolini and Barone, 2013). In general, the disadvantage is stronger for first generation immigrant students than for second generation migrants (INVALSI, 2019).

4 Data, variables and the grading bias measures

4.1 Data

Our empirical analysis is based on data provided by INVALSI, the National Institute for the Evaluation of the Italian School System. These data are particularly suitable for our purpose because they do not only provide students' performance in standardized, blindly-graded, tests (administered at the national level) in Italian, Mathematics and English[5], but also teachers' grades on the mid-term evaluation in the same subjects (derived from administrative registers) and socio-demographic student characteristics derived from both administrative registers and an ad-hoc questionnaire that is filled the day of the test. INVALSI data are collected every year on the whole population of Italian students in some

[3]During the academic year 2017/2018, the proportion of foreign students in the middle school was 10%, while it was 7,3% in upper secondary school (MIUR, 2019).

[4]However, this does not hold for the English test. In this subject, immigrants' children perform similarly to native students. They even have better performance in the listening part.

[5]The last subject, i.e. English, is present only in recent tests.

specific grades. These are grades II, V, VI, VIII, X, and XIII, although since 2013 the test is no longer administered in grade VI while the XIII test has been introduced only in the academic year 2018/2019. Every year, INVALSI collects several cross-sections of data, divided by grade and test subject. Since 2012, INVALSI data also include an individual and anonymous code which identifies every student and potentially allows to follow individuals over the educational career. For this reason, our research focuses on the period after 2012.

Our dataset is the result of a multi-step merging process of INVALSI cross-sections from different years. We merged together the Math and Italian INVALSI test datasets of grade II in the academic year 2012/2013, grade V in 2015/2016 and grade VIII in 2018/2019. We have decided to focus on the first part of the school career because we are interested in the impact of grading bias taking place at the early stages of education. Another reason is that primary and middle school are part of Italian compulsory education, which is comprehensive (students are not already sorted into different track) and still marginally characterised by dropout. Instead, the choice of the first academic year to consider, that is 2012/2013, is rather pragmatic and aimed at maximizing the number of periods observed. In fact, the cohort of students who attended grade II in 2012/2013 is the cohort that can be observed from the earlier possible grade for the longer time. Clearly, nothing prevents further research to extend our analysis to different grades and years. For instance, someone could prefer to analyse the cohort of students who attended grade V in 2013/2014 and follow it until grade X (academic year 2018/2019). This is an interesting alternative for researchers that are investigating student behaviour in a more advanced period of the school career, when students choose which track to follow in upper secondary education (see Carlana et al., 2018).

The time-invariant anonymous code allows us to track students from grade II to VIII, so that we observe students up to 3 times. Unfortunately, our dataset is not immune to attrition problems and possible selection issues, because some individuals disappear or appear over time. This fact is mainly due to two reasons: first of all, some individuals cannot be matched because they have no identification code (around 2% of our sample)[6]; secondly, some individuals with anonymous code are observed only in some of the considered grades. The latter phenomenon can be caused by a bunch of different reasons: for instance, some students may appear or disappear because they failed the admission to the following grade, because they moved from or to another country, or because they were absent the day of the INVALSI test. More information about this phenomenon can be found in Section 5.

The total number of observed students in our dataset is 619 258, but only 425 900 individuals are observed for all the three periods. Table 2 shows the number of individuals that are observed in more periods, as well as the number of students observed in only one grade.

[6]This problem mainly characterizes the academic year 2012/2013, when 1,44% of students had no identification code, while it is rather marginal in the following observed grades. In fact, in 2015/2016 only the 0,77% of observations lacked INVALSI code, while in 2018/2019 the proportion is negligible (0,28%).

4.2 Variables and descriptive statistics

The main variables included in our dataset can be distinguished between demographic variables (which describe student socioeconomic and demographic characteristics, such as gender or economic condition), institutional variables (such as school or class anonymous code, or geographical variables), student performance variables, and measures of grading bias. Some of these variables are time-invariant, while others are time-specific.

The most important demographic variables of our analysis pertain students' citizenship. We have constructed three dummy variables that distinguish between native students, first-generation foreign students (foreign-born children with at least one foreign-born parent) and second-generation foreign students (children born in Italy from foreign-born parents). The difference between first and second generation foreigners is not merely formal, but as we will see later it brings relevant implications. For the sake of simplicity, sometimes we will disregard the distinction between the two foreign status, thus focusing on the general difference between the native and foreign population. Admittedly, our dummy variables are not always capturing the different ethnicity of students because the children of immigrants who became Italian citizens before 2012 are considered as native. Our classification is therefore underestimating the incidence of immigration backgrounds. However, in our variables the information about foreign status is constructed to be time-invariant, so that missing information in a specific year is compensated by the information present in the other periods, with non-native students who obtained the Italian citizenship after 2012 that are considered as foreigners, in order to capture the different ethnicity. In our sample, the percentage of first-generation students is 3,59%, while second-generation foreign students are the 8,31%. The other socioeconomic and demographic variables, such as, for instance, gender, age, quarter of birth and socio-economic status are mainly used as controls in our models. Socio-economic status is expressed with the ESCS index[7] of the student. The ESCS index was not present in the cross-sections about grade II, and therefore in this year we exploit the value of the index in grade V, as the independence of this information from pupils' school performance in grade II can hardly be controversial.

Student performance in Math and Italian is expressed through two different realizations: teachers' grade on the midterm reporting card and score in the INVALSI test. Clearly, the midterm grade provides a non-blind evaluation, while the INVALSI assessment is blind. The comparison of the performance in these two evaluations is a key element in our analysis and is mainly carried out with respect to the performance in Math. In fact, the kind of competences that are assessed by the teacher in this subject is closer to the competences tested in the corresponding INVALSI test, while Italian teacher assessment is more prone to consider a different set of competences. As for teacher-assigned (TA) grade, the variable

[7]The ESCS is a synthetic index about the socio-economic conditions of the family of origin. It mainly considers information about parents' profession and level of education, but also the presence of some material items in the house.

ranges from 1 to 10, although grades below 5 are extremely rare[8]. The dataset contains both the written and the oral marks. We prefer to use the written marks because also INVALSI tests are written. Analogously to Triventi (2019), we exploit the oral marks to predict the missing written marks[9], since the correlation between the two types of marks is almost perfect (0,987) in grade II[10]. As it is shown in Figure 3, most of the grade distribution is around 8 and 9, which account for the 2/3 of the marks. On average, foreign students perform worse than native peers in terms of teacher's grade. Table 3 summarizes the average TA grade by year and citizenship status. In grade II, the Math average grade for natives was 8,1 and 7,4 for first-generation foreign children. Marks decreases for everyone over the years, which is evidence of stricter grading standards in the middle school, and the difference between natives' and foreigners' average grade is always relatively large and significant.

INVALSI standardized scores measure subject specific competence applying the Item Response Theory with a Rasch model to students' answer to national, blind-graded, tests. Scores are standardized to have a mean of 200 and a standard deviation of 40 in the original sample and they are later adjusted for potential cheating (Triventi, 2019). They are supposed to measure the same competences that should be developed from the curriculum taught by teachers and should therefore be comparable to what teachers' grades are expected to evaluate. We thus use the INVALSI score as proxy of student objective academic proficiency. Figure 4 describes the distribution of Italian and Math blind scores by citizenship status. The distribution for foreign students is first order stochastically dominated by that for natives, with the gap being more pronounced for first-generation foreigners.

4.3 Grading bias measures

In order to investigate the presence of systematic grading bias, we adopt an approach that has gained a foothold in the recent economic literature about this topic, that is the comparison between non-blind and blind grades (e.g. Botelho et al., 2015; Burgess and Greaves, 2013; Hanna and Linden, 2012; Lavy, 2008; Lavy and Megalokonomou, 2019; Lavy and Sand, 2018; Van Ewijk, 2011). Since the teacher-assigned grade and the INVALSI score have different ranges, they are transformed and led to a comparable and common scale. We standardize the two variables at the class level to a normal distribution so that they have the same mean (0) and standard deviation (1). If a student is performing better than the class average, she or he will therefore present a positive value, and negative otherwise. We also carry out an alternative transformation having the same scope, i.e. by ranking student performance at the

[8] In the academic year 2012/2013, only around 1% of the students had a grade below 5 in the Math and Italian grades.

[9] For instance, in grade II the percentage of students with unavailable written mark in Math is 29%. The proportion varies over the years and the subject, but it always has a relevant magnitude. Instead the percentage of students with unavailable oral mark is considerably smaller (11% of Math grades in grade II). After the manipulation, the percentage of students in grade II with unavailable written mark declines to the 6,7%.

[10] The correlation is weaker but still very close to 1 in grade V (i.e. 0,946).

class level[11]. In principle, we prefer the first manipulation because it does not only provide information about ordinality but also about cardinality. However, the ranking transformation remains an interesting and more conservative alternative for robustness checks.

Like in Alesina et al. (2018), we observe that keeping fixed the fact of belonging to a given quintile of the blind test score, foreign students receive on average a lower TA grade (see Figure 5). Native students therefore seem to benefit from a teacher premium with respect to their foreign peers, with such a difference being larger for first-generation foreign children. Similarly to Botelho at al. (2015), we therefore estimate the presence of this premium with a linear econometric model with TA grade as dependent variables and controlling for performance in the blind test. Results about grade II are shown in Table 4. Conditional on obtaining the same standardized test score, foreign students receive significantly lower grades from teachers. In fact, keeping fixed our proxy of Math competence and depending on the different model specifications, second-generation foreign children receive a grade that is between 0,279 and 0,116 standard deviations lower than native peers with similar ability. Coefficients are even larger (between -0,387 and -0,183) in the first-generation case. In our baseline model (column 1), we investigate the comparison between blind and non-blind scores standardized at the class level. Instead, the models in column 2 and 3 use as independent variable the INVALSI performance standardized at the whole population level, with the latter model that also includes class fixed-effects[12]. Results seem not to be significantly affected by the specification choice. Instead, coefficients reduce their magnitude in absolute terms when we control for a long list of other variables, that are language ability, school geographical area, sex, quarter of birth, year of birth, preschool and kindergarten attendance, delay or anticipation of the education process, and student socio-economic status. However, coefficients remain large and strongly statistically significant. We omit most of the controls' coefficients from Table 4, but their relative magnitude can be observed in Figure 6. Interestingly, natives seem not to be the only category that benefits from a grading premium, as also female students receive a grade that is slightly higher than males with similar ability[13]. However, the magnitude is significantly smaller than in the case of native status. Another category that seems to suffer from a significant gap is composed of children who were already late in their education path in grade II. This finding should not be surprising, as older age with respect to the rest of the class may potentially be perceived by teachers as a signal of lower competence, which gives birth to grading bias. Appendix Table A.1 provides the same analysis of Table 4, but here the performance variables refer to the ranking version instead of the standardization to a normal. As expected, results are very similar, although

[11] Students' performance is therefore associated to a value that ranges from 0 to 1. Individuals with the same original grade have the same ranking value.

[12] Note that class fixed-effects are not needed in the model of column 1 bucause the standardization is carried out at the class level.

[13] This finding is consistent with other results from research in education economics, such as Lavy (2008), Lavy and Megalokonomou (2019) and Lavy and Sand (2018).

the magnitude of the coefficients has decreased because the variance of the new dependent variable is smaller by construction. Appendix Tables A.2 and A.3 provide the same analyses applied to grade V in the academic year 2015/2016. Also in this case, foreign students seem to suffer a grading bias phenomenon that is large and statistically significant. Coefficients are substantially like those produced by the analysis in the previous period, even though the effect of foreign status is slightly weaker[14].

The creation of our grading bias variables follows an approach that is based on the same idea of the analyses provided above. We generate our measure by subtracting the within-class standardized measure of TA Math grade to the within-class standardized measure of student Math score in the blind test. The basic idea is therefore to measure the gap between our proxy of student objective competence and teacher assessment. A positive value means that the pupil is receiving a lower grade from the teacher with respect to a classmate with the same performance in the blind test. We also create an alternative version of this variable that instead exploits the difference between the class-rank performance of students in the blind and non-blind evaluations. The distributions of both grading bias measures have a shape that is similar to a normal, as the first is generated by the difference of two variables with normal distribution, while the second is the difference of two strongly correlated variables ranging from 0 to 1. Figure 7 shows the distribution of the grading bias measure based on the within-class standardized performance variable by citizenship status, while Figure 8 considers the rank-based alternative. In general, the variance of the standardization-based measure is larger, as the first variable has standard deviation of 0,88, while it is 0,25 in the second case[15].

Importantly, these measures are not specific to foreign students, but they rather represent a grading "gap" which, in principle, may not be due to teacher stereotypes toward a student. However, we observe that the distribution of the gap between blind and non-blind score for native students is statistically dominated by that for foreign students. We expect to find a positive relationship between our measures and foreign status, which is evidence of systematic grading bias. Table 5 shows the result of a linear regression model that investigates this fact in grade II, with the grading bias measure as dependent variable. Column 1 refers to the baseline model, while column 2 reports the coefficients after having also considered our set of controls. Coefficients are always relatively large[16], positive and statistically significant[17]. Columns 3 and 4 reports respectively the baseline and more robust model with the rank-

[14] The finding that grading bias is smaller at the end of the school cycle is consistent with the results from the existing literature, where the duration of interaction between teachers and students is negatively correlated with the bias (see also Altonji and Pierret, 2001). In fact, teachers in grade V tend to have experienced a longer interaction with students in their class than teachers in grade II.

[15] In this respect, note that the scale of the x-axis in Figure 7 and Figure 8 is different.

[16] For instance, in the baseline model first-generation students are characterised by an average grading bias that is greater than natives of 0,148, which corresponds to 0,17 standard deviations of grading bias. After having considered our controls, it reduces to 0,092, which is 0,10 standard deviations.

[17] The p-value is smaller than 0.001.

based variable alternative. Also in this case, coefficients for foreign status are positive and significant, with p-value always smaller than 0,01. In addition, we replicate the analysis in grade V (see Table 6). Coefficients are very similar to those in the previous observed period for first-generation foreigners, and generally bigger for the second-generation foreign status.

The grading bias measure that has been generated should be interpreted with caution. On average, foreign students are graded less generously by their teacher than their native peers with similar ability and similar characteristics in a broad set of other dimensions. Although we believe that our results are evidence of systematic grading bias, we cannot rule out the hypothesis that foreign status is systematically correlated with some unobserved characteristics that are considered in teacher assessment. If this assumption were confirmed, our grading bias would not only be measuring bias but also the presence of these other unobserved characteristics which eventually contribute to explain the gap between competence and teacher grade. This is probably the main weakness in our approach. On the other hand, it is encouraging the fact that when Triventi (2019) tried to explain the same grading gap with an even broader set of explanatory variables, which also includes proxies of student behavior and commitment, all the considered factors were far from accounting for the teacher grading bias against migrants' children. This suggest that implicit discrimination processes actually seem to be at work.

The determinants of the variance of grading bias are also investigated. There seems to be a negative relationship between percentage of foreign students and grading bias variance at the class and school level in both grade II and V, but in a model that controls for regional fixed effects the coefficient is not statistically different from 0 (see Appendix Tables A.4 and A.5).

In addition, we construct a further measure that presents different characteristics. We generate a variable that accounts for teacher's average grading bias by subtracting the average grading bias measure for native students at the class level to the corresponding average for foreign students. In this case, we disregard the distinction between first- and second-generation foreign students because the construction of this variable would only be effective with classes having at least one representative of every sub-population, which is not the most common case[18]. In this way, we reduce the number of generated missing values. While the individual measure of grading bias should be interpreted as the intensity of the underestimation or overestimation of subject ability that is experienced by the student, the teacher's grading bias measure provides information about the average premium that the teacher assigns to native students with respect to foreign classmates with similar ability.

Since the indicator of teacher grading bias is a variable describing the class, we investigate the correlation with other class characteristics (see Appendix Table A.6). We are concerned

[18]In academic year 2012/2013, in grade II only 1/3 of the classes had at least one student with first-generation foreign status, while around 54% of the classes had at least one student with second-generation foreign status. Instead, the percentage of classes with no foreign students is the 36% of the total.

that our measure could be capturing some class features in addition to those that we are already controlling for. Interestingly, the value of teacher bias decreases when the percentage of foreign students in the class increases. This could be explained by the fact that when in a class there are very few foreign students, these may be more prone to mechanisms of implicit stereotypes. However, this finding is not confirmed in a different model specification where teacher bias is constructed with rank variables (Appendix Table A.6, column 2), although the coefficient for the percentage of foreigners in the school remains negative and becomes significant. On the base of this, we will thus also control for class and school ethnic composition in the following analyses that use teacher bias as explanatory variables.

Now that our measures of individual grading bias and of teacher's average bias have been designed and explained, the attention will therefore be on the main research question of the present work, which pertains the estimation of the future impact of these bias phenomena on future students' academic performance.

5 Results

We now turn our attention to assessing whether children of foreign residents perform worse after having experienced under-assessment or having been taught by a biased teacher in primary school. We adopt these two different methodological and theoretical perspectives to estimate the impact of grading bias on student future outcomes. We first analyse the effect on disappearance from the sample after grade II, which is considered a negative signal of student educational attainment, although its interpretation is not univocal. Then, we analyse the impact of our grading bias metrics on medium and long-term future competence, as expressed by the performance in the blindly-graded Math INVALSI tests of grade V and VIII.

To determine the effects of our grading bias measures on future competence we estimate the following equation:

$$\begin{aligned} P_{t+s,i} = \quad & \alpha + \pi P_{t,i} + \phi F_i + + \gamma GB_{t,i} + \chi \left(F_i \cdot GB_{t,i} \right) + \\ & + \sum_{j=1}^{n} \beta_j X_{j\,t,i} + \epsilon_{t,i} \quad \forall s = 1, 2 \end{aligned}$$

where we set $t = 1$, which corresponds to grade II, while $t+1$ and $t+2$ stands for grade V and VIII, respectively. In other terms, we first estimate the equation with $s = 1$ (medium-term outcome) and in a second moment we consider $s = 2$ (long-term outcome). P stands for performance in the Math blindly-graded test; F is the dummy for foreign status; GB is the grading bias measure[19]; X is a vector of n controls and ϵ is the error term.

[19]This measure is the individual grading gap, or alternatively the average teacher grading bias, depending

In some cases, this general equation presents some minor but substantial changes. For instance, when the grading bias measure is in the version that pertains individual under-assessment rather than teacher's average bias, we also check the results from a similar equation with the two dummies F_1 (first-generation foreign status) and F_2 (second-generation foreign status) instead of the unique and general dummy variable F for foreign status. Instead, in the next subsection we investigate the case of disappearance from the sample. Therefore, in this case the dependent variable will not be blind-test performance in period $t + s$, but rather a dummy variable which takes value 1 when the individual is no longer observed after grade II.

5.1 Impact on disappearance from the sample

Before turning the attention to the effects on future blindly-graded tests, it is important to investigate the correlation between grading bias and selection into next periods sample. This is of great interest for two main reasons. The first is that disappearance from our sample is highly considered to be, on average, a negative signal of student condition, although its interpretation is not univocal. In fact, students may escape from our observation for different reasons, such as failure in the admission to the following grade, exit from the Italian school system, or simply absence in the day of the INVALSI test. We would like to estimate the impact on admission failure, but it is impossible to clearly disentangle this outcome from the others. However, the fact that, at least in statistical terms, disappearance represents a negative signal of student's condition and school attainments is mostly uncontroversial. Accordingly, students who are partially unobserved in the sample usually tend to be characterized by lower academic performance and lower socioeconomic condition in the observed period. For instance, while the average Math and Italian blind-score in grade II for student who are observed after second grade is 217 and 212, the average decreases to 202 and 197 for students that are fully unobserved after the first period. These unobserved individuals also perform worse in terms of TA grade and are more likely to be characterised by an immigration background (see Appendix Table A.7). The second point of interest in this selection analysis is that it helps us to better interpret the results about the information observed after grade II. In fact, a positive relationship between disappearance and grading bias would mean that the observed negative effects of the bias are generally underestimated.

We therefore analyse the relationship between individual under-assessment and disappearance with a linear probability model. We use our main regression equation but with a different dependent variable, that is a dummy summarizing the disappearance phenomenon. The variable takes value 1 if the individual is no longer observed after grade II and value 0 if it is observed either in grade V or VIII. In this way, we do not consider as "disappeared" the set of individuals that were absent in grade V but appeared in the next period, as well as

on the model.

students unobserved only in grade VIII. The reason is that, among the individuals who have 1 in the outcome dummy, we want to decrease the overestimation of the unknown percentage of students who failed admission before grade V. In this way, the positive value of our dependent variable can be more confidently interpreted as a negative signal of student regularity in the school path. We consider both the case of standardization-based and rank-based measures of objective performance and grading bias. Results are presented in Table 7. We observe that under-assessed students are more likely to disappear from the sample, with the effect being bigger for foreign students. As expected, the coefficients of the interaction terms between grading bias and foreign status are positive and statistically different from 0 in all the models. Appendix Table A.8 shows similar results when we distinguish between first and second generation foreign children, with the former being the more prone to disappearance and the more affected by grading bias.

The analysis is also carried out with respect to our alternative perspective about grading bias, where the measure refers to teacher's average bias in the class (see Table 8). We consider different model specifications. The baseline is a linear probability model (column 1), which also considers school fixed effects (column 2). Then we also check the results with a mixed-effect multilevel model considering the class and the school levels in the hierarchical structure. Results are quite similar in the different specifications. Even after controlling for a large set of controls (see columns 4, 5 and 6), the interaction term of foreign status and teacher bias is always positive and statistically different from 0 with p-value smaller than 0.001. Results are similar even if we consider the rank-based version of our explanatory variables (see Appendix Table A.9). We can therefore conclude that the presence of a biased teacher seems to substantially exacerbate the disappearance trend of foreign children.

5.2 Impact on medium-term future performance

The analysis now focuses on the results from our main regression equation with $s = 1$. In other terms, we estimate the impact of our grading bias metrics, which are generated in grade II, on student's Math ability in grade V, as measured by the score in the national blindly-graded test. In our model, individual under-assessment takes place in the first observed period, and we assess to what extent the intensity of this "treatment" on foreign students leads to decreasing the math score in the second period, compared to peers with similar characteristics. We keep Math competence fixed in grade II using the blind-test score standardized at the national level, while the grading bias measure exploits the comparison between blind and non-blind grade at the class level. The outcome variable is Math score in grade V standardized at the national level, considering only students who were matched between the first and the second period[20]. We use a linear regression model with standard

[20]However, the standardization at the national level considering also new students observed in grade V and not in grade II produces extremely similar results.

errors clustered at the class level and results are also put to the test of our typical set of controls (see Table 9, columns 1 and 2). Finally, similar models with the rank-based version of math score and grading bias are considered (columns 3 and 4)[21].

From the results of our regressions, we observe that under-assessed students are characterised by poorer future performance with respect to individuals with same competence in grade II. This holds for both native and foreign students, which reminds us to be cautious in the interpretation of grading bias coefficients. For every unitary increase in grading bias, the relative performance decreases between 0,19 and 0,15 standard deviations. In terms of the ranking-based model, this is equivalent to saying that for every unitary increase in the alternative grading bias measure, the ranking of the student at the national level decreases from 15 to 18 centiles in a scale from 0 to 100. However, the effect turns out to be stronger for foreign students. The children of foreign residents suffer more teacher under-assessment than native students by a 14%[22] in the baseline model (column 1), which raises to a 19% when considering more controls (column 2). Similarly, in the ranking-based models the interaction term coefficients are the 13% of the grading bias one. Interaction terms are always characterized by negative and significant coefficients.

The outcomes from the regression models that distinguish between foreign subcategories can be consulted in Appendix Table A.10. According to these results, first generation foreign students suffer more teacher under-assessment than native students by a 12% in the baseline model, which raises to a 24% when considering more controls. This trend is confirmed in the ranking-based models, with the interaction term coefficients being between the 9% and the 16% of the grading bias one. Interaction terms are always characterized by negative and significant coefficients, apart from the baseline rank-based model (column 3). Instead, interaction terms' coefficients are always negative and statistically different from 0 in the case of second-generation foreign students. The magnitude and the interpretation are relatively similar to the case of first-generation foreign status, although here the size of the effect seems to slightly decrease by adding controls, while it was the contrary for the other foreign category[23].

In a nutshell, we can say that ceteris paribus, our grading bias measure is strongly and negatively correlated with medium-term future school performance, although caution is needed in the interpretation. The significant and negative coefficients of the interaction terms with foreign status seem to confirm that grading bias have a negative impact on the

[21] Importantly, note that the ranking of the Math score at the national level ranges from 0 to 100. This choice is aimed at allowing a more intuitive interpretation of the coefficients. Instead the ranking measure that were used in the grading bias variable are ranked in the interval from 0 to 1.

[22] In fact, the coefficient of the grading bias variable is -0,190 in the baseline model, with the interaction term for foreign students being -0,027.

[23] In addition to the medium-term effects of grade II bias on grade V performance, we have also carried out a similar analysis with grade V as first period and grade VIII as second (see Appendix Table A.11). Results are similar to the comparison between grade II and V, although in the ranking-based version models interaction terms are not statistically different from 0.

future attainments of students with immigration background. Nonetheless, the magnitude of the grading bias coefficients in Table 9, who characterize also native students, reminds that if the grading bias is taken alone, it only provides a measure of individual under-assessment which is not always due to ethnic bias.

The interpretation of our estimates is instead relatively easier in the case of teacher bias effect. Teacher bias measure is capturing the difference in the blind vs. non-blind score average gap between foreign and native students, and therefore it provides a measure of how much foreign children are underassessed with respect to their objective competence, after having subtracted the average under assessment of native class-mates. Now, we estimate and comment the impact of teacher bias in grade II on medium-term future competence, that is performance in the blind-test of grade V. Results are reported in Table 10. Similarly to Table 8, the different model specifications include a linear regression model with clustered standard errors at the class level, plus a model with also school fixed effects, and a mixed-effect multilevel model. We then consider the same models with the addition of our broad set of controls, which also includes class characteristics such as number of students and percentage of foreign children. The positive coefficient associated to teacher bias can be interpreted as a premium that benefits native students and lead them to perform better in the future with respect to native peers who had no premium. The magnitude of the effects of teacher bias on foreign students is, instead, much bigger and associated to a negative phenomenon. The coefficients of the interaction terms between -0,149 and -0,214 mean that for every standard deviation increase in teacher bias (+ 0,74), the future performance of foreign students decreases by a value between 0,10 and 0,14 standard deviations[24]. The coefficients of the interaction terms are always statistically different from zero with a p-value below 0,001. The results from the ranking-based approach are reported in the Appendix Table A.12 and show similar findings. In this case, since the standard deviation of teacher's bias is 0,2, it can be inferred that, for every standard deviation increase in the bias, the ranking of future performance decreases by a number of centiles that is between 3 and 4[25].

The bottom line is thus that, based on our results, having a biased teacher may very well lead foreign students to poorer performance, as opposed to foreign students with similar characteristics and less biased teachers. Moreover, the findings from the sample disappearance suggests that the negative effects found in this section are likely to be underestimated.

Finally, we also provide a prima facie analysis of the relationship between the variance of the individual grading bias measure in the class (or school) and future average class (or school) performance. Results can be consulted in Appendix Tables A.13 and A.14. We observe a negative relationship between these two variables. Although this relationship is interesting to the scope of our analysis, it is left to further research, as the focus will remain

[24] This is equivalent to saying that for every unitary increase in teacher bias foreign students have a score in grade V test that is between 0,137 and 0,190 sd lower than foreign peers with teacher bias equal to 0.

[25] These values correspond to 1/5 of 15 and 20, the absolute value of the smaller and bigger interaction term coefficients.

on individual effects.

5.3 Impact on long-term future performance

After having investigated the effects of grading bias in the second year of primary school on Math competence three years later, we turn our attention to the farthest period available, that is the last year of the middle school. In formal terms, the main regression equation is applied considering $s = 2$, i.e. exploiting the information available in the first period to explain the Math score in the third observed period. The empirical strategy is equivalent to the estimation of medium-term effects, with the only difference being the period of the dependent variable.

Like in the previous analysis, we first estimate the future effects of grading bias considering our measure of the gap between competence and TA grade, and then we adopt the approach based on teacher average bias in the class. The results of the former analysis can be found in Table 11 and Appendix Table A.15, while the outputs regarding the latter are presented in Table 12 and Appendix Table A.16.

In the previous section, we found that individual under-assessment leads to future weaker performance for both native and foreign students, with the effect being stronger for the latter. Table 11 shows that this is also true when we estimate the impact on performance in grade VIII. However, here the negative component that is common to native and foreign students is relatively smaller, while the detrimental effect that is specific to non-natives increases. As expected, all interaction terms are negative and statistically different from 0. In the baseline model (column 1), the marginal effect of a unitary increase[26] in grading bias to foreign students corresponds to an incremental decrease in long-term future performance in Math blind test of 0,052 standard deviations. The magnitude decreases to 0,034 after controlling for our typical set of additional independent variables. In general, it can be observed (see Appendix Table A.15) that in all the specifications the negative impact of grading bias is stronger for first-generation foreign students, with the exemption of the last model (column 4) where the coefficient of children with first-generation migration background significantly shrinks[27].

As expected, also with the approach based on teacher average bias we detect significant and negative impact of the phenomenon under analysis (see Table 12). The negative effects of grading bias are probably underestimated because individuals with lower initial competence, a foreign background, and proneness to suffer from grading bias are more likely to disappear

[26] In the case of grading bias in the standardization-based form, that has standard deviation equal to 0,88, a unitary increase corresponds to an increase of around 1,13 standard deviations.

[27] In the Appendix Table A.15 all the interaction terms are negative and significant,with the exemption of first-generation foreign interaction in the ranking-based model with several controls (column 4). In the baseline model (column 1), the marginal effect of a unitary increase in grading bias to first-generation and second-generation foreign students corresponds to an incremental decrease in long-term future performance in Math blind test of respectively 0,057 and 0,049 standard deviations. The magnitude decreases to 0,038 and 0,033 after controlling for our typical set controls.

from the sample after grade II. Ceteris paribus, for every unitary increase of teacher bias (which corresponds to 1,35 sd of the independent variable), the drop in foreigners' long-term future performance in the blind-graded test is between 0,093 and 0,210 standard deviations. The estimated effect generally decreases after the introduction of several controls (columns 4, 5 and 6), it is bigger in the mixed-effect multilevel model that considers classes and schools in the hierarchical structure (column 3), and it is smaller in the baseline model (columns 1 and 4). The main findings are substantially confirmed when we replicate the analysis with the ranking-based alternative of the models (see Appendix Table A.16).

6 Conclusion

In this study, we have tried to shed light on a specific mechanism that may contribute to exacerbate and settle the pattern of persistent educational inequality between native and foreign students. The case of Italy is extremely interesting because the relatively recent increase of the non-native student population constitutes a strategic but also new structural challenge, testing the capacity of the national education system to provide equality of opportunity.

We have found evidence of systematic grading bias against foreign students. In general, students with an immigration background receive lower teacher-assigned grades than natives, after their performance on anonymously graded standardized tests is held constant. On the base of this, we have constructed some individual and class measures of relative under-assessment of foreign students, which operate as an indicator of grading bias. Unfortunately, this work is not immune to the typical fundamental identification problem of the economic literature about grading bias, which is the possible presence of additional unobserved factors explaining the gap between blind and non-blind grades. However, despite not being perfect, our measures remain a reliable approximation and are consistent with the existing literature.

Therefore, we have evaluated the impact of being affected by teacher grading bias on future school performance. Foreign students that are under-assessed in their intra-classroom evaluation or that are taught by an averagely biased teacher seem to be characterised by a more likely irregular education path and perform worse in future blindly-graded assessment compared to both foreign and native peers with similar characteristics but with no experience of significant teacher bias.

Despite its limitations, our study pointed out an overlooked source of educational inequalities that can negatively influence the integration of foreign pupils in the country. Our research also involves some relevant policy implications. In fact, an intervention that can potentially bring positive spillovers could be inspired at increasing the awareness of principals and teachers about how grading standards may unintentionally change between native and foreign students. This can be done by providing simple and effective reports or, as suggested by Alesina et al. (2018), by encouraging every teacher to take an Implicit Association

Test that provides feedbacks about the presence of both explicit and implicit stereotypes. Moreover, as proposed by Triventi (2019) and in line with previous recommendations by educational psychologists (Malouff, 2008; Malouff et al., 2013), introducing students anonymity during teacher grading might contribute to mitigating social inequalities in scholastic assessment, which reflect not only student competence in the subject but also socially biased perceptions.

To conclude, our research provides further evidence that discrimination may lead to self-fulfilling prophecies by negatively influencing the future performance of discriminated groups. This finding highlights a complex and key challenge for the education systems in western countries. Education is widely considered to be the key determinant to foster children of immigrants' socioeconomic integration in host countries, which lead to a positive impact on the whole society. However, policymakers and academics should be aware that this process is still slowed down by the mechanisms that fossilize education inequality.

References

Ackermann, K. and Schlicht-Schm, R. (2012). Logics of Educational Stratification: A Cross-National Map of Educational Inequality, *International Scholarly Research Network.*

Akifyeva, R. and Alieva, A. (2016). The Influence of Ethnicity on Teacher Expectations and Teacher Perceptions of Student Warmth and Competence Basic Research Program. *National Research University Higher School of Economics, Working papers series: Psychology*, No. WP BRP 65/PSY/2016, Moscow.

Altonji, J. and Pierret, C. (2001). Employer Learning and Statistical Discrimination. *The Quartely Journal of Economics*, Vol.116. No. 1 (Feb.2001), pp. 313-350.

Alesina A., Carlana M., La Ferrara E., Pinotti P. (2018). Revealing Stereotypes: Evidence from immigrants in schools, *National Bureau of Economic Research.*

Arrow, K. (1973). The Theory of Discrimination. Princeton University Press, 3-33. Princeton, N.J.

Azzolini, D. (2011). A New Form of Educational Inequality? What We Know and What We Still Do Not Know About the Immigrant-Native Gap in Italian Schools. *Italian Journal of Sociology of Education* 3(1):197–222.

Azzolini, D. and Barone, C. (2013). Do They Progress Or Do They Lag Behind? Educational Attainment of Immigrants' Children in Italy: The Role Played by Generational Status, Country of Origin and Social Class. *Research in Social Stratification and Mobility* 31:82–96.

Barbieri, G., Rossetti, C., and Sestito, P. (2011). The determinants of teacher mobility: Evidence using Italian teachers' transfer applications. *Economics of Education Review*, 30(6):1430–1444.

Bordalo, P., Coffman, K., Gennaioli, N. and Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics.*

Botelho, F., Madeira, R. A., and Rangel, M. A. (2015). Racial discrimination in grading: Evidence from Brazil. *American Economic Journal: Applied Economics*, 7(4):37–52.

Burgess, S. and Greaves, E. (2013). Test scores, subjective assessment, and stereotyping of ethnic minorities. *Journal of Labor Economics*, 31(3):535–576.

Barban, N., and White, M. J. (2011). Immigrants' Children's Transition to Secondary School in Italy. *International Migration Review* 45(3):702–26.

Carlana, M., La Ferrara, E., and Pinotti, P. (2018). Goals and gaps: Educational careers of immigrant children. Mimeo Bocconi University.

Casteel, C. A. (1998), Teacher-Student Interactions and Race in Integrated Classrooms, *Journal of Educational Research*, 92, 115–120.

Coate, S., and Loury, G., C. (1993). Will affirmative-action policies eliminate negative stereotypes?, *American Economic Review*, 83(5), 1220-1240.

Dee, T. S. (2005), A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?, *American Economic Review*, 95(2),158–65.

Hoff, K., and Pandey, P. (2006). Discrimination, social identity, and durable inequalities. *American Economic Review*, 96(2), 206-211.

Gilliam, W. S., Maupin, A. N., Reyes, C. R., Accavitti, M., and Shic, F. (2016). Do early educators' implicit biases regarding sex and race relate to behavior expectations and recommendations of preschool expulsions and suspensions. Research Study Brief. Yale University.

Glover, D., Pallais, A., and Parienté, W. (2017). Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores. *The Quarterly Journal of Economics.*

Grusky, D. B. and Hauser, R. M. (1984). Comparative social-mobility revisited—models of convergence and divergence in 16 countries, *American Sociological Review.*

Hanna, R. N. and Linden, L. L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy*, 4(4):146–68.

Good, T.L. (1987), Two Decades of Research on Teacher Expectations: Findings and Future Directions, *Journal of Teacher Education*, 38(4), 32-47.

INVALSI, (2019). Rilevazione Nazionale Degli Apprendimenti 2018–2019: Rapporto Risultati. *INVALSI.* Roma.

ISTAT (2020), Bilancio demografico nazionale, *ISTAT*. Roma.

Hinnerich, B. Höglin, and E., Johannesson, M. (2015). Discrimination against students with foreign backgrounds: evidence from grading in Swedish public high schools. *Education Economics.*

Jussim, L. and Harber, K. D. (2005). Teacher Expectations and Self-Fulfilling Prophecies: Knowns Unknowns, Resolved and Unresolved Controversies. *Personality and Social Psychology Review*, 9(2), 131–155.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics*, 92(10):2083–2105.

Lavy, V. and Megalokonomou , R. (2019). Persistency in Teachers' Grading Bias and Effects on Longer-Term Outcomes: University Admissions Exams and Choice of Field of Study. *NBER Working Paper.* No. 26021.

Lavy, V. and Sand, E. (2018). On the origins of gender human capital gaps: Short and long term consequences of teachers' stereotypical biases. *Journal of Public Economics.*

Lindahl, E. (2007). Comparing teachers' assessments and national test results - evidence from Sweden. Working Paper Series 2007:24, IFAU - Institute for Evaluation of Labour Market and Education Policy.

Lundberg, S. J., and Startz, R. (1983). Private discrimination and social intervention in competitive labor market, *American Economic Review*, 73(3), 340-347.

Malouff, J. M. (2008). Bias in Grading. *College Teaching* 56:3, 191–92.

Malouff, J. M., Emmerton A. J., and Schutte, N. S. (2013). The Risk of a Halo Bias as a Reason to Keep Students Anonymous During Grading. *Teaching of Psychology* 40:233–37.

Mechtenberg, L. (2006). Cheap Talk in the Classroom: How biased grading at school explains gender differences in achievements, career choices, and wages. *Review of Economic Studies.*

MIUR (2019). Gli alunni con cittadinanza non italiana - Anno scolastico 2017/2018, *MIUR* Roma.

Rosenthal, R. and Jacobson, L. (1968). Pygmalion in the classroom: Teacher expectations and student intellectual development. New York: Holt, Rinehart and Winston.

Schlicht-Schmalzle, R., and Ackermann, K. (2012). Logics of Educational Stratification: A Cross-National Map of Educational Inequality, *International Scholarly Research Network.*

Sewell, W. H. and Shah, V. P. (1967). Socioeconomic status, intelligence, and attainment of higher education. *Sociology of Education.*

Sprietsma, M. (2013). Discrimination in Grading: Experimental Evidence from Primary School Teachers. *Empirical Economics* 45(1):1–16.

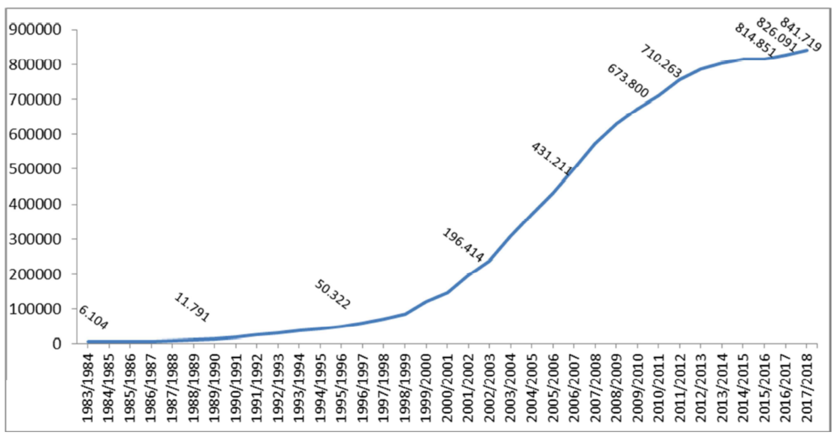
Triventi, M. (2019). Are Children of Immigrants Graded Less Generously by their Teachers than Natives, and Why? Evidence from Student Population Data in Italy. *International Migration Review.*

Van Ewijk, R. (2011). Same Work, Lower Grade? Student Ethnicity and Teachers' Subjective Assessments. *Economics of Education Review* 30(5):1045–58.

Figures and Tables*

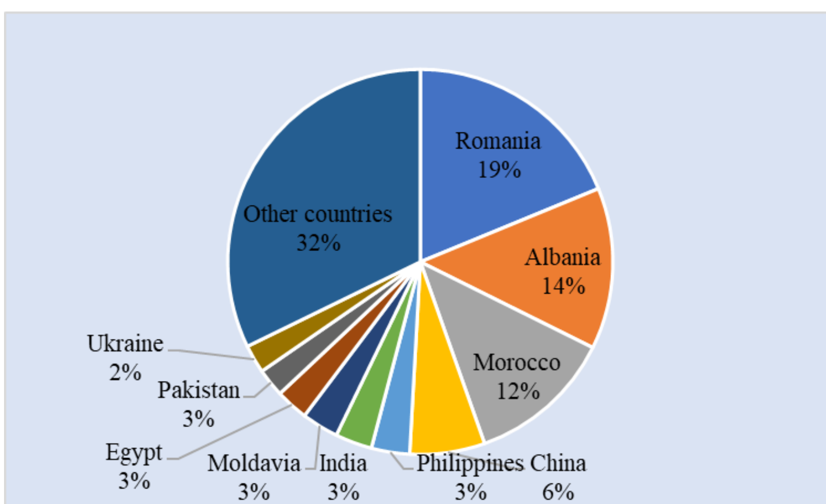
*General note: In all the regression tables the p-value of the coefficients is expressed with the following significance stars: * for $p < .05$, ** for $p < .01$, and *** for $p < .001$.
When not otherwise specified, the adopted set of controls is made of language ability, school geographic area, sex, quarter of birth, year of birth, preschool and kindergarten attendance, delay or anticipation of the education process, and student socio-economic status.

Figure 1: Number of students with non-Italian citizenship from 1983 to 2018



Source: Italian Ministry of Education (MIUR - "Ufficio Statistica e studi")

Figure 2: Nationality of immigrant students



Source: Italian Ministry of Education (MIUR - "Ufficio Statistica e studi")

Figure 3: Non-blind grade frequency and box plots by citizenship status in Grade II (2012)

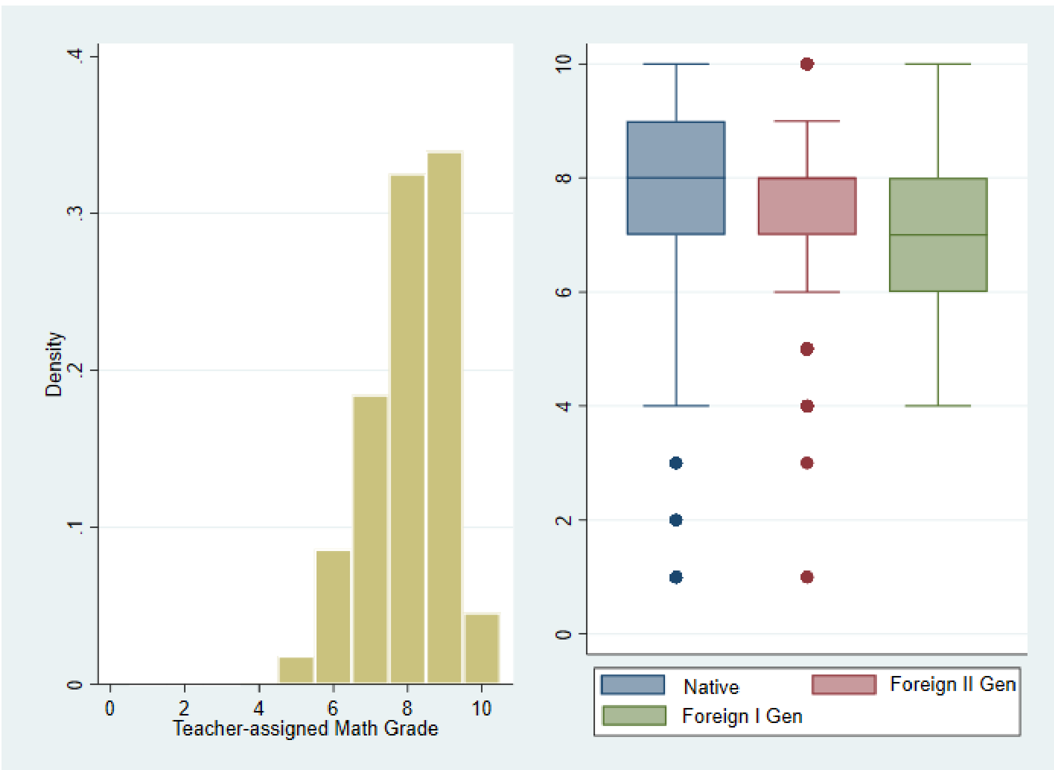


Figure 4: Blind score distribution by citizenship status in Grade II (year 2012)

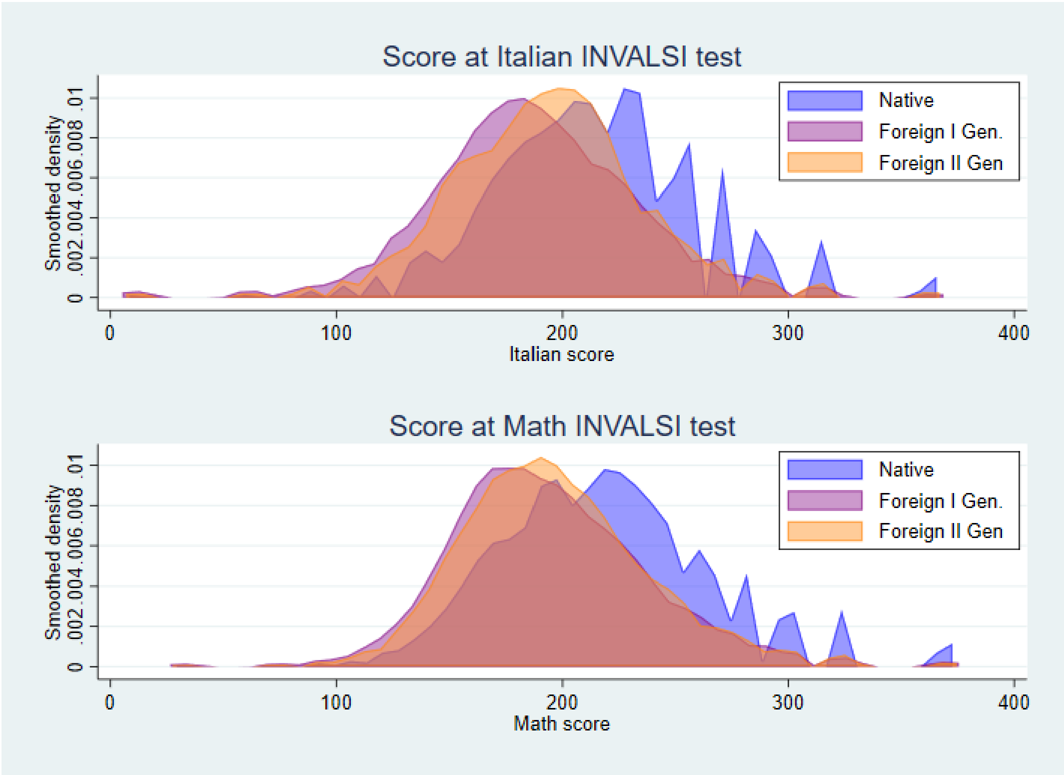


Figure 5: Class TA Rank by citizenship status and Blind Rank quintile

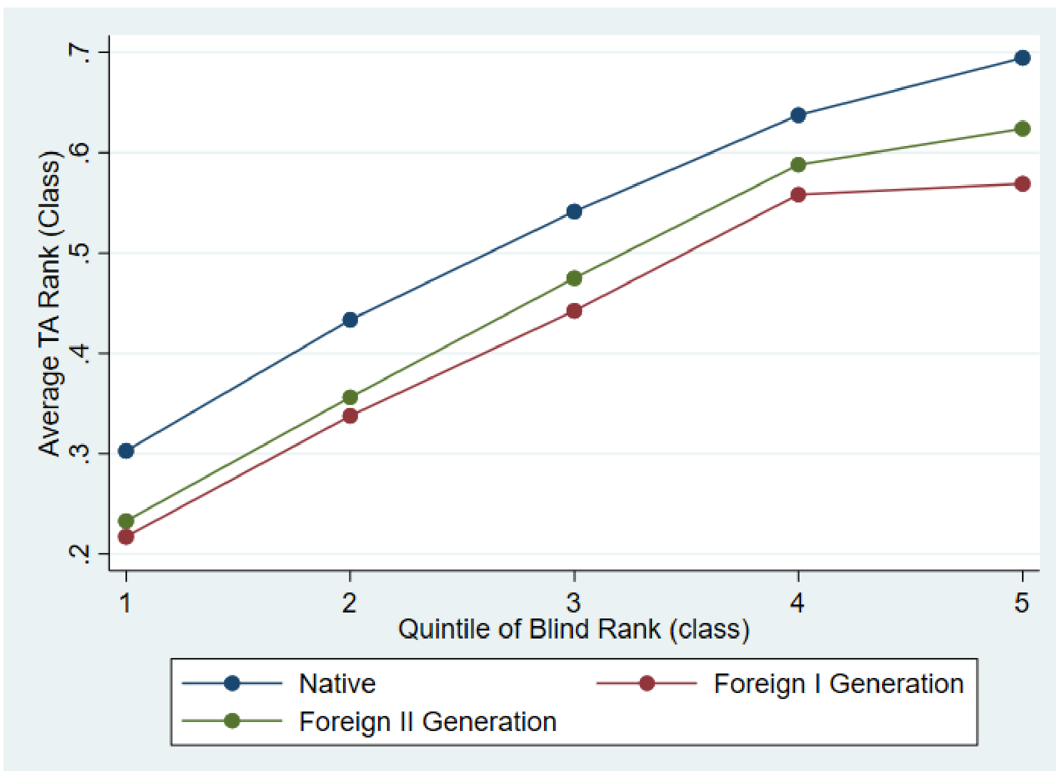
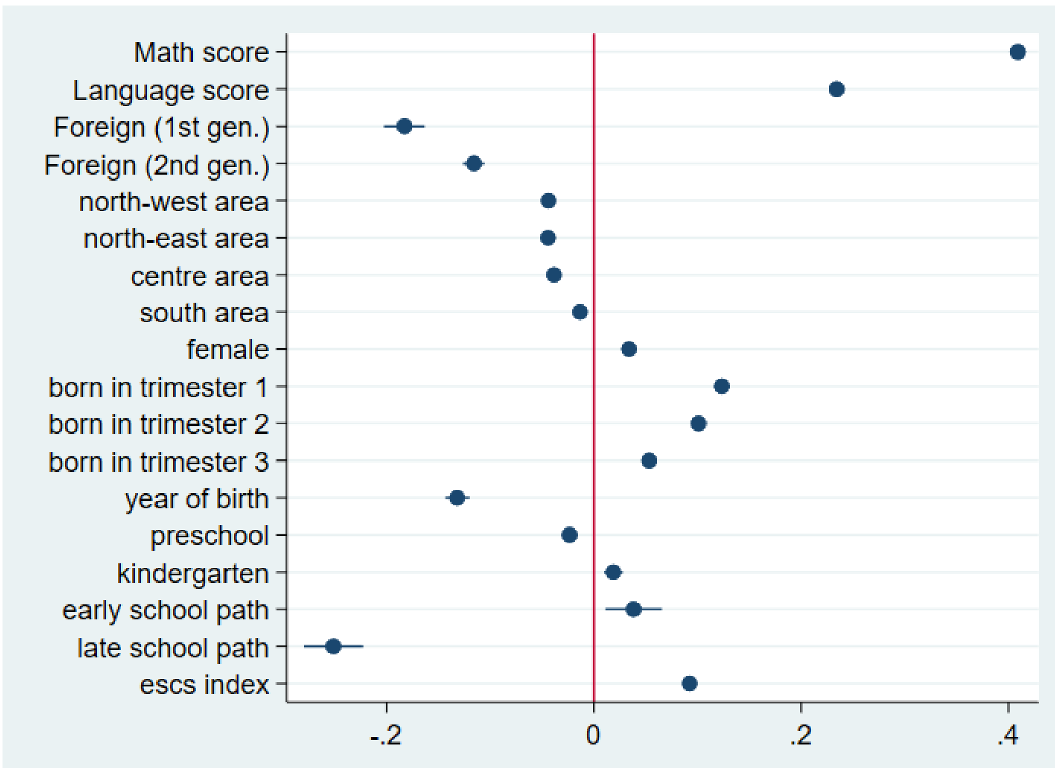


Figure 6: Coefficients of grading bias regression



Note: These are the coefficients from a regression with grading bias in grade II as dependent variable. Math and Language scores represent the performance in the INVALSI test standardized at class level. The baseline for the area's dummy is the area of south and islands, which includes Basilicata, Calabria, Sicily and Sardinia. The baseline for the dummies about the quarter of birth refers to birth in the last trimester. The escs index is the individual socio-economic background in 2015.

Figure 7: Grading bias (Z) distribution

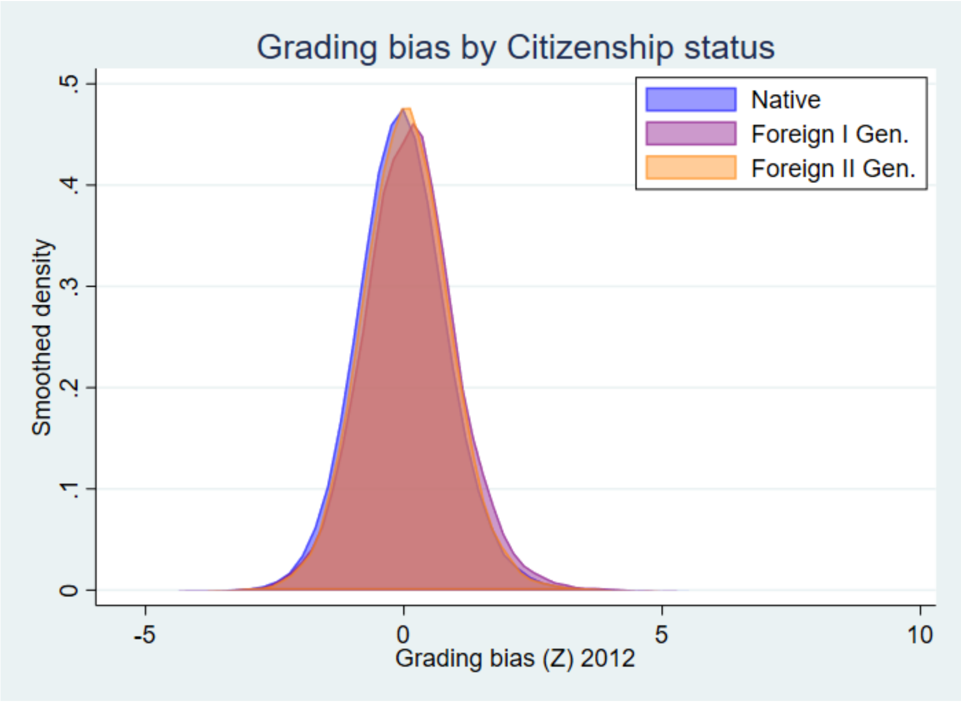


Figure 8: Grading bias (Rank) distribution

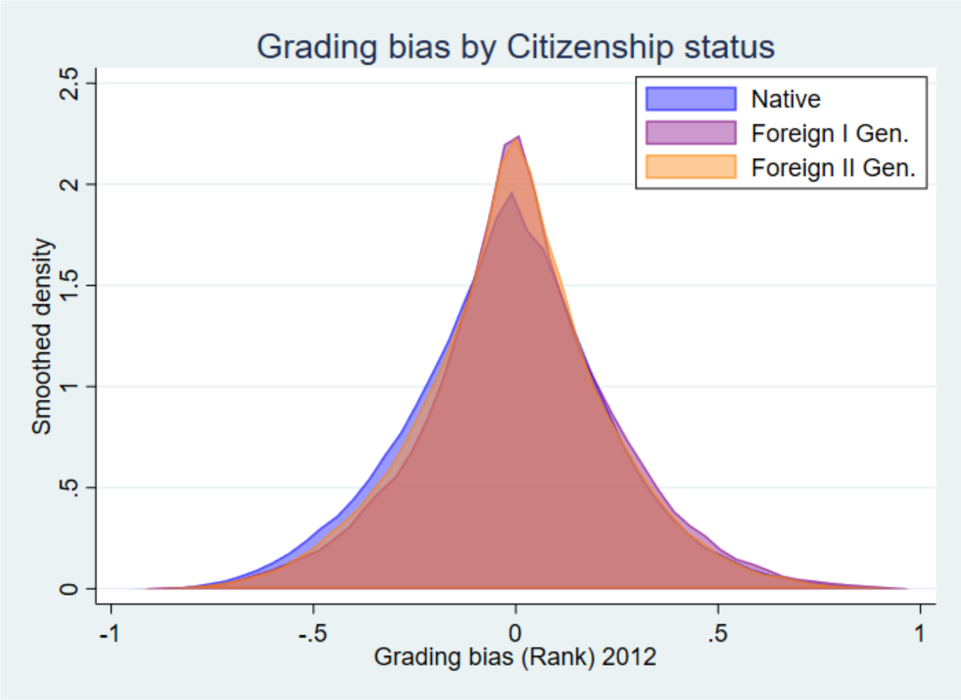


Table 1: Country of birth of immigrant students (school year 2017-2018)

Country of birth	Number of students by school level					Share
	Kinderg.	Primary	Second. (I)	Second. (II)	Total	
Romania	31 850	57 408	31 943	36 843	158 044	19%
Albania	23 412	40 966	23 406	26 456	114 240	14%
Morocco	23 050	42 180	19 951	18 035	103 216	12%
China	9 795	20 042	13 413	10 089	53 339	6%
Philippines	3 941	8 383	6 362	8 376	27 062	3%
India	5 875	10 756	4 975	4 867	26 473	3%
Moldavia	4 657	8 262	5 024	7 603	25 546	3%
Egypt	5 273	9 091	4 791	3 740	22 895	3%
Pakistan	3 751	8 598	4 415	3 660	20 424	2%
Ukraine	3 138	6 376	4 190	6 183	19 887	2%
Other countries	50 373	95 756	55 345	69 119	270 593	32%
Foreign students	165 115	307 818	173 815	194 971	841 719	100%
	19,6%	36,6	20,7	23,2	100%	100%

Source: Italian Ministry of Education (MIUR - "Ufficio Statistica e studi")
The table reports the total number of students by country of birth and by school level in year 2017-2018 for the most represented nationalities, and their share among all immigrant students.

Table 2: Attrition in the dataset

Matched individuals by grade			
	Grade II	Grade V	Grade VIII
Grade II	517 778	451 961	457 235
Grade V		496 285	457 930
Grade VIII			546 411
Individuals observed in only one grade			
Grade II		34 482	
Grade V		12 294	
Grade VIII		57 146	
Total number of observations		619 258	
Observations observed in all grades		425 900	

Table 3: Average teacher-assigned grades

	Grade II		Grade V		Grade VIII	
	Math	Italian	Math	Italian	Math	Italian
Native	8.08	7.98	7.92	7.87	6.87	7.07
Foreign II Generation	7.57	7.34	7.44	7.29	6.49	6.66
Foreign I Generation	7.36	7.05	7.20	7.01	6.27	6.52

Table 4: Estimation of grading bias against foreign students in grade II (2012)

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. variable:			**Grade given by Math teacher (Z)**			
Foreign (1st gen.)	-0.339***	-0.383***	-0.387***	-0.183***	-0.235***	-0.211***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Foreign (2nd gen.)	-0.228***	-0.245***	-0.279***	-0.116***	-0.152***	-0.146***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
Math score (Z-class)	0.573***			0.409***		
	(0.00)			(0.00)		
Math score (Z)		0.416***	0.737***		0.286***	0.520***
		(0.00)	(0.00)		(0.00)	(0.00)
Set of controls				√	√	√
Constant	0.039***	0.041***	0.043***	0.533***	0.804***	0.529***
	(0.00)	(0.00)	(0.00)	(0.02)	(0.03)	(0.03)
R-sqr	0.350	0.205	0.355	0.396	0.243	0.415
Observations	464 862	465 005	465 005	329 864	330 048	330 048

Note: The dependent variable is standardized to a normal at the class level. The models in (1) and (4) uses as control the math score standardized at the class level, whereas the models in (2), (3), (5) and (6) exploit the score standardized in the whole population. The model in (3) and (6) includes class fixed effects. The models in (4), (5) and (6) contain our set of controls.

Table 5: Determinants of grading bias in grade II

	(1)	(2)	(3)	(4)
Dep. variable:	**Grading bias (Z)**		**Grading bias (Rank)**	
Foreign (1st gen.)	0.148***	0.092***	0.029***	0.018***
	(0.01)	(0.01)	(0.00)	(0.00)
Foreign (2nd gen.)	0.072***	0.034***	0.014***	0.005**
	(0.00)	(0.01)	(0.00)	(0.00)
Set of controls		$\checkmark$		$\checkmark$
Constant	-0.020***	-0.076**	-0.019***	-0.037***
	(0.00)	(0.03)	(0.00)	(0.01)
R-sqr	0.001	0.006	0.001	0.005
Observations	464 862	340 396	466 702	341 701

Table 6: Determinants of grading bias in grade V

	(1)	(2)	(3)	(4)
Dep. variable:	**Grading bias (Z)**		**Grading bias (Rank)**	
Foreign (1st gen.)	0.150***	0.090***	0.033***	0.020***
	(0.01)	(0.01)	(0.00)	(0.00)
Foreign (2nd gen.)	0.098***	0.067***	0.022***	0.014***
	(0.00)	(0.00)	(0.00)	(0.00)
Set of controls		$\checkmark$		$\checkmark$
Constant	-0.023***	-0.065**	-0.015***	-0.030***
	(0.00)	(0.02)	(0.00)	(0.01)
R-sqr	0.002	0.016	0.001	0.013
Observations	462 450	388 839	463 193	389 441

Table 7: Effects of grading bias on disappearance from the sample

	(1)	(2)	(3)	(4)
Dependent variable:	**Disappearance after grade II**			
	Standardized measures		Rank measures	
Math score	-0.018***	-0.010***	-0.001***	-0.000***
	(0.00)	(0.00)	(0.00)	(0.00)
Grading bias	0.017***	0.011***	0.045***	0.029***
	(0.00)	(0.00)	(0.00)	(0.00)
Foreign	0.129***	0.093***	0.131***	0.094***
	(0.00)	(0.00)	(0.00)	(0.00)
Foreign x	0.012***	0.010***	0.026***	0.018*
Grading bias	(0.00)	(0.00)	(0.00)	(0.01)
Set of controls		√		√
Constant	0.047***	0.063***	0.076***	0.084***
	(0.00)	(0.01)	(0.00)	(0.01)
R-sqr	0.040	0.030	0.038	0.029
Observations	464 862	375 955	466 702	377 404

Note: The dependent variable is a dummy variable taking value 1 if the individual is no longer observed after grade II, and taking value 0 if it is observed either in grade V or VIII. All the explanatory variables refer to grade II. Math score is expression of student's relative performance at the whole population level, while grading bias is generated at the class level. Both variables are in the standardization-based version in (1) and (2), and in the rank-based one in (3) and (4). The models in (2) and (4) also include our typical set of controls.

Table 8: Effects of teacher bias on disappearance from the sample

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. variable:	**Disappearance after grade II**					
Math score	-0.018***	-0.024***	-0.025***	-0.008***	-0.010***	-0.011***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Teacher bias	-0.003*	-0.002**	-0.003**	-0.002	-0.000	-0.002
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Foreign	0.138***	0.135***	0.133***	0.096***	0.094***	0.094***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Foreign x	0.026***	0.028***	0.030***	0.019***	0.019***	0.020***
Teacher bias	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Set of controls				√	√	√
Constant	0.044***	0.044***	0.046***	0.108***	0.080***	0.121***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
R-sqr	0.052	0.204		0.040	0.242	
Observations	301 188	301 188	301 188	245 210	245 210	245 210

Note: The dependent variable is a dummy variable taking value 1 if the individual is no longer observed after grade II, and taking value 0 if it is observed either in grade V or VIII. All the explanatory variables refer to grade II, with Math score and grading bias in the standardization-based version. (1) presents the baseline model with s.e. clustered at the class level. In (2) we also include school fixed effects. In (3) we use a multi-level mixed effects model considering both the class and school levels. (4), (5) and (6) present the same models with our set of controls. We also control for number of students in the class and percentage of foreigners.

Table 9: Impact of grading bias on medium-term future performance

	(1)	(2)	(3)	(4)
Dep. variable:	**Math score 2015**			
	Standardized measure		Rank measure	
Math score 2012	0.605***	0.497***	0.617***	0.517***
	(0.00)	(0.01)	(0.00)	(0.00)
Grading bias	-0.190***	-0.152***	-18.475***	-15.289***
	(0.00)	(0.00)	(0.20)	(0.22)
Foreign	-0.091***	0.002	-2.403***	0.217
	(0.00)	(0.01)	(0.13)	(0.18)
Foreign x	-0.027***	-0.029***	-2.168***	-2.060***
Grading bias	(0.01)	(0.01)	(0.54)	(0.62)
Set of controls		√		√
Constant	-0.013***	-0.129***	18.472***	13.739***
	(0.00)	(0.04)	(0.08)	(1.05)
R-sqr	0.317	0.342	0.334	0.360
Observations	301 188	398 457	329 132	329 132

Note: In (1) and (2) the dependent variable is the score in the Math blind test in grade V and it is standardized at the whole population level considering only individuals that are matched between grade II and V. Instead in (3) and (4) the Math score is ranked at the whole population level (still considering individuals that are matched between grade II and V only) and takes a value from 0 to 100. All explanatory variables refer to grade II, with Math score and grading bias in the standardization-based version in (1) and (2), while it is in the ranking-based version in (3) and (4). (1) and (3) present the baseline model with s.e. clustered at the class level. (2) and (4) present the same models with our set of controls.

Table 10: Impact of teacher bias on medium-term future performance

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. variable:			**Math score 2015**			
Math score 2012	0.589***	0.611***	0.637***	0.471***	0.498***	0.519***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Teacher bias	0.019**	0.030***	0.024***	0.012	0.023***	0.016*
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Foreign	-0.098***	-0.074***	-0.058***	0.004	0.014**	0.019***
	(0.01)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Foreign x	-0.181***	-0.192***	-0.214***	-0.149***	-0.162***	-0.175***
Teacher bias	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Set of controls				√	√	√
Constant	-0.013**	-0.017***	-0.023***	-0.121*	-0.205***	-0.257***
	(0.00)	(0.00)	(0.00)	(0.05)	(0.04)	(0.04)
R-sqr	0.330	0.494		0.364	0.521	
Observations	258 242	258 242	258 242	214 966	214 966	214 966

Note: The dependent variable is the score in the Math blind test in grade V and it is standardized at the whole population level considering only individuals that are matched between grade II and V. All the explanatory variables refer to grade II, with Math score and grading bias in the standardization-based version. (1) presents the baseline model with s.e. clustered at the class level. In (2) we also include school fixed effects. In (3) we use a multi-level mixed effects model considering both the class and school levels. (4), (5) and (6) present the same models with our set of controls. We also control for number of students in the class and percentage of foreigners.

Table 11: Impact of grading bias on long-term future performance

	(1)	(2)	(3)	(4)
Dep. variable:	**Math score 2018**			
	Standardized measure		Rank measure	
Math score 2012	0.460***	0.383***	0.494***	0.406***
	(0.00)	(0.00)	(0.00)	(0.00)
Grading bias 2012	-0.122***	-0.086***	-13.001***	-10.305***
	(0.00)	(0.00)	(0.18)	(0.22)
Foreign	0.038***	0.033***	1.617***	1.420***
	(0.01)	(0.01)	(0.15)	(0.18)
Foreign x	-0.052***	-0.034***	-3.670***	-1.354*
Grading bias 2012	(0.01)	(0.01)	(0.61)	(0.68)
Set of controls		$\surd$		$\surd$
constant	-0.021***	-0.568***	24.412***	5.350***
	(0.00)	(0.03)	(0.09)	(0.96)
R-sqr	0.181	0.327	0.210	0.352
Observations	413 889	314 177	415 509	315 388

Note: In (1) and (2) the dependent variable is the score in the Math blind test in grade VIII and it is standardized at the whole population level considering only individuals that are matched between grade II and VIII. Instead in (3) and (4) the Math score is ranked at the whole population level (still considering individuals that are matched between grade II and VIII only) and takes a value from 0 to 100. All the explanatory variables refer to grade II, with Math score and grading bias in the standardization-based version in (1) and (2), while it is in the ranking-based version in (3) and (4). (1) and (3) present the baseline model with s.e. clustered at the class level. (2) and (4) present the same models with our set of controls.

Table 12: Impact of teacher bias on long-term future performance

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. variable:			**Math score 2018**			
Math score 2012	0.496***	0.649***	0.711***	0.405***	0.522***	0.572***
	(0.01)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Teacher bias	0.005	0.021***	0.019**	0.014**	0.012**	0.025***
	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)
Foreign	-0.066***	-0.058***	-0.020***	0.040***	0.073***	0.089***
	(0.01)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
Foreign x	-0.148***	-0.195***	-0.230***	-0.107***	-0.144***	-0.169***
Teacher bias	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Set of controls				$\surd$	$\surd$	$\surd$
Constant	0.091***	0.091***	0.063***	-0.748***	-0.497***	-1.212***
	(0.00)	(0.00)	(0.00)	(0.05)	(0.04)	(0.04)
R-sqr	0.224	0.409		0.342	0.452	
Observations	264 117	264 117	264 117	203 834	203 834	203 834

Note: The dependent variable is the score in the Math blind test in grade VIII and it is standardized at the whole population level considering only individuals that are matched between grade II and VIII. All the explanatory variables refer to grade II, with Math score and grading bias in the standardization-based version. (1) presents the baseline model with s.e. clustered at the class level. In (2) we also include school fixed effects. In (3) we use a multi-level mixed effects model considering both the class and school levels. (4), (5) and (6) present the same models with our set of controls. We also control for number of students in the class and percentage of foreigners.

Appendix Tables*

*General note: In all the regression tables the p-value of the coefficients is expressed with the following significance stars: * for p < .05, ** for p < .01, and *** for p < .001.
When not otherwise specified, the adopted set of controls is made of language ability, school geographic area, sex, quarter of birth, year of birth, preschool and kindergarten attendance, delay or anticipation of the education process, and student socio-economic status.

Appendix Table A.1: Estimation of grading bias against foreign students in grade II (2012)

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. variable:	**Grade given by Math teacher (Rank)**					
Foreign (1st gen.)	-0.086***	-0.090***	-0.094***	-0.045***	-0.055***	-0.050***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Foreign (2nd gen.)	-0.061***	-0.059***	-0.068***	-0.030***	-0.036***	-0.035***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Math score (R.-class)	0.543***			0.398***		
	(0.00)			(0.00)		
Math score (Rank)		0.004***	0.007***		0.003***	0.005***
		(0.00)	(0.00)		(0.00)	(0.00)
Set of controls				√	√	√
Constant	0.258***	0.326***	0.199***	0.363***	0.516***	0.316***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
R-sqr	0.349	0.231	0.370	0.397	0.265	0.427
Observations	466 702	466 702	466 702	331 220	331 220	331 220

Note: The dependent variable is the score ranking at the class level. The model in (1) and (4) uses as control the math score ranking at the class level, whereas the model in (2), (3), (5) and (6) exploits the score ranking in the whole population. The model in (3) and (6) includes class fixed effects. The models in (4), (5) and (6) contain our set of controls.

Appendix Table A.2: Estimation of grading bias against foreign students in grade V (2015)

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. variable:	**Grade given by Math teacher (Z)**					
Foreign (1st gen.)	-0.308***	-0.359***	-0.345***	-0.103***	-0.131***	-0.107***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Foreign (2nd gen.)	-0.205***	-0.222***	-0.245***	-0.075***	-0.095***	-0.083***
	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.01)
Math score (Z-class)	0.624***			0.428***		
	(0.00)			(0.00)		
Math score (Z)		0.449***	0.812***		0.263***	0.553***
		(0.00)	(0.00)		(0.00)	(0.00)
Set of controls				√	√	√
Constant	0.036***	0.040***	0.042***	0.203***	0.329***	0.171***
	(0.00)	(0.00)	(0.00)	(0.02)	(0.03)	(0.02)
R-sqr	0.408	0.232	0.411	0.483	0.307	0.492
Observations	462 450	462 535	462 535	378 918	379 031	379 031

Note: The dependent variable is standardized at the class level. The model in (1) and (4) uses as control the math score standardized at the class level, whereas the models in (2), (3), (5) and (6) exploit the score standardized in the whole population. The model in (3) and (6) includes class fixed effects. The models in (4), (5) and (6) contain our typical set of controls.

Appendix Table A.3: Estimation of grading bias against foreign students in grade V (2015)

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. variable:			**Grade given by Math teacher (Rank)**			
Foreign (1st gen.)	-0.081***	-0.090***	-0.090***	-0.028***	-0.034***	-0.029***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Foreign (2nd gen.)	-0.055***	-0.056***	-0.063***	-0.019***	-0.023***	-0.020***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Math score (R.-class)	0.592***			0.409***		
	(0.00)			(0.00)		
Math score (Rank)		0.004***	0.008***		0.003***	0.005***
		(0.00)	(0.00)		(0.00)	(0.00)
Set of controls				√	√	√
Constant	0.230***	0.313***	0.162***	0.230***	0.356***	0.167***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
R-sqr	0.403	0.249	0.419	0.482	0.322	0.500
Observations	463 193	463 193	463 193	379 541	379 541	379 541

Note: The dependent variable is the score ranking at the class level. The model in (1) and (4) uses as control the math score ranking at the class level, whereas the models (2), (3), (5) and (6) exploit the score ranking in the whole population. The model in (3) and (6) includes class fixed effects. The models in (4), (5) and (6) contain our typical set of controls.

Appendix Table A.4: Grading bias variance at the class level in grade II and V

	(1)	(2)	(3)	(4)
Dependent variable:	**Bias variance (2012)**		**Bias variance (2015)**	
Perc. of foreign (2012)	-0.044	-0.004		
	(0.03)	(0.00)		
Number of students (2015)	-0.002**	-0.001***		
	(0.00)	(0.00)		
Perc. of foreign (2015)			-0.020	-0.004
			(0.02)	(0.00)
Number of students (2015)			0.000	-0.000***
			(0.00)	(0.00)
ESCS school (2015)	0.066***	0.004***	0.052***	0.004***
	(0.01)	(0.00)	(0.01)	(0.00)
Constant	0.875***	0.078***	0.729***	0.067***
	(0.02)	(0.00)	(0.02)	(0.00)
R-sqr	0.064	0.064	0.046	0.045
Observations	26 224	26 469	27 148	27 294

Note: In (1) and (3) the dependent variable, i.e. grading bias, is constructed with standardized versions of students' performance, whereas in (2) and (4) grading bias is originated from the comparison of ranking variables. In this analysis the observations are classes and not students.

Appendix Table A.5: Grading bias variance at the school level in grade II and V

	(1)	(2)	(3)	(4)
Dependent variable:	**Bias variance (2012)**		**Bias variance (2015)**	
Perc. of foreign (2012)	-0.041	-0.002		
	(0.04)	(0.00)		
Number of students (2012)	-0.001***	-0.000***		
	(0.00)	(0.00)		
Perc. of foreign (2015)			0.024	0.000
			(0.04)	(0.00)
Number of students (2015)			-0.001***	-0.000***
			(0.00)	(0.00)
ESCS school (2015)	0.094***	0.004***	0.082***	0.005***
	(0.01)	(0.00)	(0.01)	(0.00)
Constant	0.911***	0.070***	0.805***	0.065***
	(0.01)	(0.00)	(0.01)	(0.00)
R-sqr	0.149	0.134	0.132	0.131
Observations	6 685	6 749	6 671	6 687

Note: In (1) and (3) the dependent variable, i.e. grading bias, is constructed with standardized versions of students' performance, whereas in (2) and (4) grading bias is originated from the comparison of ranking variables. In this analysis the observations are schools and not students.

Appendix Table A.6: Determinants of average teacher grading bias in grade II

	(1)	(2)
Dependent variable:	**Teacher bias (Z)**	**Teacher bias (Rank)**
Number of students (class)	0.0006	0.0002*
	(0.00)	(0.00)
Number of students (school)	0.0003***	0.0001***
	(0.00)	(0.00)
Perc. of foreigners (class)	-0.0895***	0.0107**
	(0.02)	(0.00)
Perc. of foreigners (school)	-0.0367	-0.0326***
	(0.02)	(0.01)
Constant	0.0969***	0.0116***
	(0.01)	(0.00)
R-sqr	0.001	0.000
Observations	311 561	312 293

Note: In (1) the dependent variable is the main teacher grading bias version, constructed using the standardized measures of grading bias, while the dependent variable in (2) comes from the rank-based alternatives

Appendix Table A.7: Characteristics of observed and unobserved individuals after grade II

	(1)	(2)
Subpopulation:	**Observed**	**Unobserved**
Math score 2012	216.51	202.14
Language score 2012	212.42	196.56
TA Math grade	8.06	7.36
TA Language grade	7.95	7.21
Percentage of natives	.90	.67
Perc. of foreign (1st gen.)	.02	.12
Perc. of foreign (2nd gen.)	.07	0.21
Observations	482 945	32 544

Note: (1) refers to the population of students observed in grade II and later observed also in either grade V or VIII. Instead, (2) describes the population of individuals who, after being observed in grade II, disappear in the next periods. The Math and Language score in the blind test are in the original scale. Teacher-assigned grades are the written marks in the midterm reporting card.

Appendix Table A.8: Effects of grading bias on disappearance from the sample

	(1)	(2)	(3)	(4)
Dependent variable:	**Disappearance after grade II**			
	Standardized measures		Rank measures	
Math score	-0.018***	-0.010***	-0.001***	-0.000***
	(0.00)	(0.00)	(0.00)	(0.00)
Grading bias	0.017***	0.011***	0.045***	0.029***
	(0.00)	(0.00)	(0.00)	(0.00)
Foreign (1st gen.)	0.193***	0.127***	0.196***	0.128***
	(0.00)	(0.00)	(0.00)	(0.00)
Foreign (1st gen.) x	0.011***	0.013**	0.027**	0.031
Grading bias	(0.00)	(0.00)	(0.01)	(0.02)
Foreign (2nd gen.)	0.106***	0.082***	0.108***	0.083***
	(0.00)	(0.00)	(0.00)	(0.00)
Foreign (2nd gen.) x	0.010***	0.008**	0.011***	0.011
Grading bias	(0.00)	(0.00)	(0.01)	(0.01)
Set of controls		√		√
Constant	0.047***	0.060***	0.076***	0.081***
	(0.00)	(0.01)	(0.00)	(0.01)
R-sqr	0.043	0.031	0.041	0.029
Observations	464 862	375 955	466 702	377 404

Note: The dependent variable is a dummy variable taking value 1 if the individual is no longer observed after grade II, and taking value 0 if it is observed either in grade V or VIII. All explanatory variables refer to grade II. Math score is expression of student's relative performance at the whole population level, while grading bias is generated at the class level. Both variables are in the standardization-based version in (1) and (2), and in the rank-based one in (3) and (4). The models in (2) and (4) also include our typical set of controls.

Appendix Table A.9: Effects of teacher bias (Rank) on disappearance from the sample

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. variable:	**Disappearance after grade II**					
Math score	-0.001***	-0.001***	-0.001***	-0.000***	-0.000***	-0.000***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Teacher Bias	-0.005	-0.004	-0.005	-0.002	0.001	-0.001
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Foreign	0.139***	0.137***	0.135***	0.097***	0.096***	0.095***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Foreign x	0.074***	0.077***	0.084***	0.050***	0.049***	0.054***
Teach. bias	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Set of controls				√	√	√
Constant	0.073***	0.081***	0.084***	0.133***	0.110***	0.151***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
R-sqr	0.051	0.202		0.040	0.241	
Observations	301 896	301 896	301 896	245 756	245 756	245 756

Note: The dependent variable is a dummy variable taking value 1 if the individual is no longer observed after grade II, and taking value 0 if it is observed either in grade V or VIII. All explanatory variables refer to grade II, with Math score and grading bias in the ranking-based version. (1) presents the baseline model with s.e. clustered at the class level. In (2) we also include school fixed effects. In (3) we use a multi-level mixed effects model considering both the class and school levels. (4), (5) and (6) present the same models with our set of controls. We also control for number of students in the class and percentage of foreigners.

Appendix Table A.10: Impact of grading bias on medium-term future performance

	(1)	(2)	(3)	(4)
Dep. variable:	**Math score 2015**			
	Standardized measure		Rank measure	
Math score 2012	0.605***	0.497***	0.617***	0.517***
	(0.00)	(0.01)	(0.00)	(0.00)
Grading bias	-0.190***	-0.152***	-18.475***	-15.289***
	(0.00)	(0.00)	(0.20)	(0.22)
Foreign (1st gen.)	-0.099***	0.019	-2.767***	0.516
	(0.01)	(0.01)	(0.29)	(0.32)
Foreign (1st gen.) x	-0.023*	-0.036**	-1.824	-2.500*
Grading bias	(0.01)	(0.01)	(1.13)	(1.23)
Foreign (2nd gen.)	-0.088***	-0.003	-2.292***	0.136
	(0.01)	(0.01)	(0.19)	(0.20)
Foreign (2nd gen.) x	-0.029***	-0.027***	-2.248***	-1.950**
Grading bias	(0.01)	(0.01)	(0.65)	(0.70)
Set of controls		√		√
Constant	-0.013***	-0.130***	18.474***	13.717***
	(0.00)	(0.04)	(0.19)	(1.05)
R-sqr	0.317	0.342	0.334	0.360
Observations	301 188	398 457	329 132	329 132

Note: In (1) and (2) the dependent variable is the score in the Math blind test in grade V and it is standardized at the whole population level considering only individuals that are matched between grade II and V. Instead in (3) and (4) the Math score is ranked at the whole population level (still considering individuals that are matched between grade II and V only) and takes a value from 0 to 100. All the explanatory variables refer to grade II, with Math score and grading bias in the standardization-based version in (1) and (2), while it is in the ranking-based version in (3) and (4). (1) and (3) present the baseline model with s.e. clustered at the class level. (2) and (4) present the same models with our set of controls.

Appendix Table A.11: Impact of grading bias (grade V) on medium-term future performance (grade VIII)

	(1)	(2)	(3)	(4)
Dep. variable:	**Math score 2018**			
	Standardized measures		Rank measures	
Math score 2015	0.538***	0.372***	0.555***	0.377***
	(0.00)	(0.00)	(0.00)	(0.00)
Grading bias	-0.147***	-0.087***	-14.689***	-9.531***
	(0.00)	(0.00)	(0.17)	(0.21)
Foreign (1st gen.)	-0.074***	0.047***	-2.238***	1.148***
	(0.01)	(0.01)	(0.25)	(0.27)
Foreign (1st gen.) x	-0.039***	-0.024*	-1.701	1.159
Grading bias 2015	(0.01)	(0.01)	(1.07)	(1.17)
Foreign (2nd gen.)	0.055***	0.043***	1.740***	1.336***
	(0.01)	(0.01)	(0.15)	(0.17)
Foreign (2nd gen.) x	-0.017*	-0.019**	-0.056	0.476
Grading bias 2015	(0.01)	(0.01)	(0.66)	(0.69)
Set of controls		√		√
Constant	-0.016***	-0.404***	21.313***	5.387***
	(0.00)	(0.03)	(0.08)	(0.80)
R-sqr	0.255	0.419	0.274	0.436
Observations	429 535	357 058	430 193	357 592

Note: In (1) and (2) the dependent variable is the score in the Math blind test in grade VIII and it is standardized at the whole population level considering only individuals that are matched between grade V and VIII. Instead in (3) and (4) the Math score is ranked at the whole population level (still considering individuals that are matched between grade V and VIII only) and takes a value from 0 to 100. All the explanatory variables refer to grade V, with Math score and grading bias in the standardization-based version in (1) and (2), while it is in the ranking-based version in (3) and (4). (1) and (3) present the baseline model with s.e. clustered at the class level. (2) and (4) present the same models with our set of controls. We also control for number of students in the class and percentage of foreigners.

Appendix table A.12: Impact of teacher bias on medium-term future performance (Rank)

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. variable:			**Math score 2015**			
Math score 2012	0.600***	0.618***	0.638***	0.489***	0.510***	0.528***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Teacher bias	2.019**	2.608***	2.569***	1.170	1.974***	1.747**
	(0.65)	(0.53)	(0.62)	(0.66)	(0.55)	(0.66)
Foreign	-2.660***	-2.040***	-1.681***	0.146	0.413**	0.533***
	(0.16)	(0.13)	(0.11)	(0.15)	(0.14)	(0.13)
Foreign x	-17.849***	-18.851***	-20.914***	-14.966***	-16.165***	-17.360***
Teacher bias	(0.94)	(0.79)	(0.66)	(0.99)	(0.86)	(0.72)
Set of controls				√	√	√
Constant	19.559***	18.536***	17.356***	14.417***	10.598***	8.406***
	(0.20)	(0.15)	(0.15)	(1.49)	(1.16)	(1.13)
R-sqr	0.352	0.494		0.383	0.519	
Observations	258 836	258 836	258 836	215 435	215 435	215 435

Note: The dependent variable is the score in the Math blind test in grade V and it is ranked at the whole population level (between 0 and 100) considering only individuals that are matched between grade II and V. All the explanatory variables refer to grade II, with Math score and grading bias in the ranking-based version. (1) presents the baseline model with s.e. clustered at the class level. In (2) we also include school fixed effects. In (3) we use a multi-level mixed effects model considering both the class and school levels. (4), (5) and (6) present the same models with our set of controls. We also control for number of students in the class and percentage of foreigners.

Appendix table A.13: Effect of grading bias variance on future average class performance

	(1)	(2)	(3)	(4)
Dependent variable:	**Avg. score (2015)**		**Avg. score (2018)**	
Bias variance (Z) (2012)	-2.747***			
	(0.45)			
Bias variance (Rank) (2012)		-29.958***		
		(5.53)		
Bias variance (Z) (2015)			-3.260***	
			(0.23)	
Bias variance (Rank) (2015)				-35.634***
				(2.93)
Avg. math score (2012)	0.405***	0.399***		
	(0.01)	(0.01)		
Perc. of foreign (2012)	-4.480**	-4.382**		
	(1.40)	(1.40)		
Number of students (2012)	0.144***	0.129**		
	(0.04)	(0.04)		
Avg. math score (2015)			0.122***	0.121***
			(0.00)	(0.00)
Perc. of foreign (2015)			-9.947***	-10.314***
			(0.81)	(0.81)
Number of students (2015)			0.204***	0.185***
			(0.02)	(0.02)
ESCS school (2015)	5.354***	5.306***	11.939***	11.918***
	(0.71)	(0.71)	(0.31)	(0.31)
Constant	125.146***	126.248***	175.037***	175.376***
	(2.12)	(2.13)	(0.94)	(0.96)
R-sqr	0.218	0.216	0.508	0.504
Observations	26 151	26 395	27 079	27 223

Note: In this analysis the observations are classes and not students. The dependent variable is average math score in the class in the blindly-graded test of grade V ((1) and (2)) and VIII ((3) and (4)). Also the independent variables refer to the class, when not otherwise specified. (1) and (3) exploit the standardization based variables, while (2) and (4) use the rank-based alternatives.

Appendix table A.14: Effect of grading bias variance on future average school performance

	(1)	(2)	(3)	(4)
Dependent variable:	**Avg. score (2015)**		**Avg. score (2018)**	
Bias Variance (Z) (2012)	-3.613**			
	(1.21)			
Bias Variance (Rank) (2012)		-33.533*		
		(15.13)		
Bias Variance (Z) 2015			-4.317***	
			(0.57)	
Bias Variance (Rank) (2015)				-48.423***
				(7.48)
Avg. math score (2012)	0.436***	0.431***		
	(0.02)	(0.02)		
Perc. of foreign (2012)	2.467	2.161		
	(2.56)	(2.56)		
Number of students (2012)	0.007	0.008		
	(0.01)	(0.01)		
Avg. math score (2015)			0.066***	0.062***
			(0.01)	(0.01)
Perc. of foreign (2015)			-5.126***	-5.166***
			(1.33)	(1.34)
Number of students (2015)			0.027***	0.028***
			(0.00)	(0.00)
ESCS school (2015)	6.119***	5.864***	12.876***	12.828***
	(0.77)	(0.77)	(0.32)	(0.32)
Constant	120.542***	120.858***	187.773***	188.262***
	(3.29)	(3.27)	(1.24)	(1.26)
R-sqr	0.277	0.277	0.654	0.651
Observations	6 684	6 748	6 654	6 669

Note: In this analysis the observations are schools and not students. The dependent variable is average math score in the school in the blindly-graded test of grade V ((1) and (2)) and VIII ((3) and (4)). Also the independent variables refer to the school. (1) and (3) exploit the standardization based variables, while (2) and (4) use the rank-based alternatives.

Appendix Table A.15: Impact of grading bias on long-term future performance

	(1)	(2)	(3)	(4)
Dep. variable:	**Math score 2018**			
	Standardized measure		Rank measure	
Math score 2012	0.460***	0.383***	0.494***	0.406***
	(0.00)	(0.00)	(0.00)	(0.00)
Grading bias 2012	-0.122***	-0.086***	-12.994***	-10.305***
	(0.00)	(0.00)	(0.18)	(0.22)
Foreign (1st gen.)	-0.045***	0.044***	-0.833**	1.654***
	(0.01)	(0.01)	(0.30)	(0.35)
Foreign (1st gen.) x	-0.057***	-0.038**	-3.740**	-0.806
Grading bias 2012	(0.01)	(0.01)	(1.21)	(1.45)
Foreign (2nd gen.)	0.063***	0.030***	2.344***	1.356***
	(0.01)	(0.01)	(0.17)	(0.19)
Foreign (2nd gen.) x	-0.049***	-0.033***	-3.478***	-1.524*
Grading bias 2012	(0.01)	(0.01)	(0.69)	(0.75)
Set of controls		√		√
constant	-0.021***	-0.568***	24.421***	5.336***
	(0.00)	(0.03)	(0.09)	(0.96)
R-sqr	0.181	0.327	0.211	0.352
Observations	413 889	314 177	415 509	315 388

Note: In (1) and (2) the dependent variable is the score in the Math blind test in grade VIII and it is standardized at the whole population level considering only individuals that are matched between grade II and VIII. Instead in (3) and (4) the Math score is ranked at the whole population level (still considering individuals that are matched between grade II and VIII only) and takes a value from 0 to 100. All the explanatory variables refer to grade II, with Math score and grading bias in the standardization-based version in (1) and (2), while it is in the ranking-based version in (3) and (4). (1) and (3) present the baseline model with s.e. clustered at the class level. (2) and (4) present the same models with our set of controls.

Appendix table A.16: Impact of teacher bias on long-term future performance (Rank)

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. variable:			**Math score 2018**			
Math score 2012	0.510***	0.627***	0.674***	0.409***	0.499***	0.539***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Teacher bias	-0.605	1.692***	0.882	1.059*	1.008*	2.104**
	(0.61)	(0.42)	(0.64)	(0.51)	(0.43)	(0.68)
Foreign	-1.579***	-1.659***	-0.667***	1.375***	2.128***	2.530***
	(0.16)	(0.14)	(0.13)	(0.17)	(0.16)	(0.15)
Foreign x	-14.619***	-19.241***	-22.319***	-10.934***	-14.691***	-17.110***
Teacher bias	(0.95)	(0.84)	(0.78)	(1.03)	(0.95)	(0.87)
Set of controls				√	√	√
Constant	27.007***	21.142***	18.014***	-0.387	1.437	-20.212***
	(0.15)	(0.12)	(0.16)	(1.35)	(1.17)	(1.24)
R-sqr	0.250	0.421		0.363	0.461	
Observations	264 744	264 744	264 744	204 286	204 286	204 286

Note: The dependent variable is the score in the Math blind test in grade VIII and it is ranked at the whole population level (between 0 and 100) considering only individuals that are matched between grade II and VIII. All the explanatory variables refer to grade II, with Math score and grading bias in the ranking-based version. (1) presents the baseline model with s.e. clustered at the class level. In (2) we also include school fixed effects. In (3) we use a multi-level mixed effects model considering both the class and school levels. (4), (5) and (6) present the same models with our set of controls. We also control for number of students in the class and percentage of foreigners.